Journal of Hydrology 571 (2019) 765–781

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Formulating a strategy to combine artificial intelligence models using Bayesian model averaging to study a distressed aquifer with sparse data availability

Marjan Moazamnia[a], Yousef Hassanzadeh[a], Ata Allah Nadiri[b], Rahman Khatibi[c], Sina Sadeghfam[d,*]

[a] Faculty of Civil Engineering, University of Tabriz, Tabriz, East Azerbaijan, Iran
[b] Department of Earth Sciences, Faculty of Natural Sciences, University of Tabriz, Tabriz, East Azerbaijan, Iran
[c] GTEV-ReX Limited, Swindon, Wilts, United Kingdom
[d] Department of Civil Engineering, Faculty of Engineering, University of Maragheh, Maragheh, East Azerbaijan, Iran

ARTICLE INFO

This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Bayesian model averaging
Distressed Urmia aquifer
Management plans

ABSTRACT

A modelling strategy is formulated, which collectively consists of separate Multiple Models (MM) and uses Bayesian Model Averaging (BMA) to combine these MMs to learn from data. The procedure is at two levels: at Level 1, three Artificial Intelligence (AI) models are constructed, which comprise Artificial Neural Network (ANN), Sugeno Fuzzy Logic (SFL) and Multiple-Neuro-Fuzzy (Multi-NF) but their outputs are directed to BMA at the next level; at Level 2, BMA is used to combine ANN, SFL and Multi-NF for better predictions and with facilities for quantifying uncertainty. The model performance is tested using the data from Urmia aquifer in the West Azerbaijan province, northwest Iran, where due to the absence of participatory water usage management practices both Lake Urmia and its surrounding 12 aquifers (including the study area) are distressed. The modelling strategy and its results on Urmia aquifer provides an insight to the study area and will be used to investigate ways of arresting the decline in the water table of the aquifer but this should be feasible only by developing a series of management plans, including basin management plans, aquifer management plans, drought plans, water cycle studies. Under such an integrated management system, the model developed here is demonstrably well-placed to serve as an operational management tool for the aquifer.

1. Introduction

Bayesian Model Averaging (BMA), transformed into a practical tool since Draper (1995) and Höting et al. (1999), is a strategy to combine Multiple Models (MM) often constructed by perturbing parameters; and to use its capability for assessing inherent uncertainties. This paper investigates performances of BMA by combining separate MMs comprising three different Artificial Intelligence (AI) techniques for predicting Groundwater Level (GWL) of Urmia aquifer. The strategy is notable for coping with sparse data, as the case is for Urmia aquifer, and for providing an insight into the state of the aquifer and its declining GWLs during the 14 years of its record. The study touches on the broader context of the problem, where Lake Urmia is in a distressed state since 2000 due to overambitious embankment dams and this is paralleled by over-abstractions of its aquifers. Fig. 1 compares the decline in Lake Urmia water level (a decline of 0.3 m/year) with declining GWL in Urmia aquifer (an approx. decline of 0.2 m/year), where both declines are attributable to the absence of any policy, or poor policies at both national and regional levels. The paper aims to gain an insight into the problem as the economic vibrancy of the population at the fertile Urmia plain depends on maintaining the integrity of the high-quality of the aquifer.

Model combination is a topical research and includes: (i) ensemble modelling often using MMs by perturbing parameters; see Clemen (1989) for a comprehensive review and Cloke and Pappenberger, (2009) for a review on meteorological forecasting; (ii) applications of BMA in various disciplines since the works by Draper (1995) and as elaborated by Höting et al. (1999); and (iii) the authors' research activities in hydrology and geohydrology often use an AI model to combine MMs (AIMM), e.g. see Nadiri et al. 2017, Nadiri et al. 2018a. These

* Corresponding author.
E-mail addresses: m.moazzamniya@tabrizu.ac.ir (M. Moazamnia), yhassanzadeh@tabrizu.ac.ir (Y. Hassanzadeh), nadiri@tabrizu.ac.ir (A.A. Nadiri), gtev.rex@gmail.com (R. Khatibi), s.sadeghfam@maragheh.ac.ir (S. Sadeghfam).

https://doi.org/10.1016/j.jhydrol.2019.02.011
Received 8 November 2018; Received in revised form 29 January 2019; Accepted 1 February 2019
Available online 16 February 2019
0022-1694/

Nomenclature

AI	artificial intelligence
AIC	Akaike information criterion
ANFIS	adaptive neuro-fuzzy inference system
ANN	artificial neural network
BIC	Bayesian information criterion
BMA	Bayesian model averaging
C_Δ	variance matrix of prediction errors
D	observed data (GWL)
FL	fuzzy logic
GA	genetic algorithm
GEP	gene expression programming
GTRAP	input datasets which is acronym for groundwater level, temperature, river discharge, abstraction, precipitation
GTRP	input datasets without abstraction
GWL	groundwater level
HBMA	hierarchical Bayesian model averaging
KIC	Kashyap information criterion
LM	levenberg-marquardt
MF	membership function
MCS	Monte Carlo simulation
MFL	Mamdani fuzzy logic
MLP	multi-layer perceptron
MMs	multiple models
m_p	number of model parameter
OK	ordinary Kriging
OW	observation well
PDF	probability density function
Q_p	sum of the weighted squared errors
R^2	determination coefficient
RMSE	root mean square error
SC	subtractive clustering
SFL	Sugeno fuzzy logic
SOM	self organizing map
SVM	support vector machines
WAWA	west Azerbaijan regional water authority
A	scaling factor
Δ	predicted quantity
λ	parsimony parameter

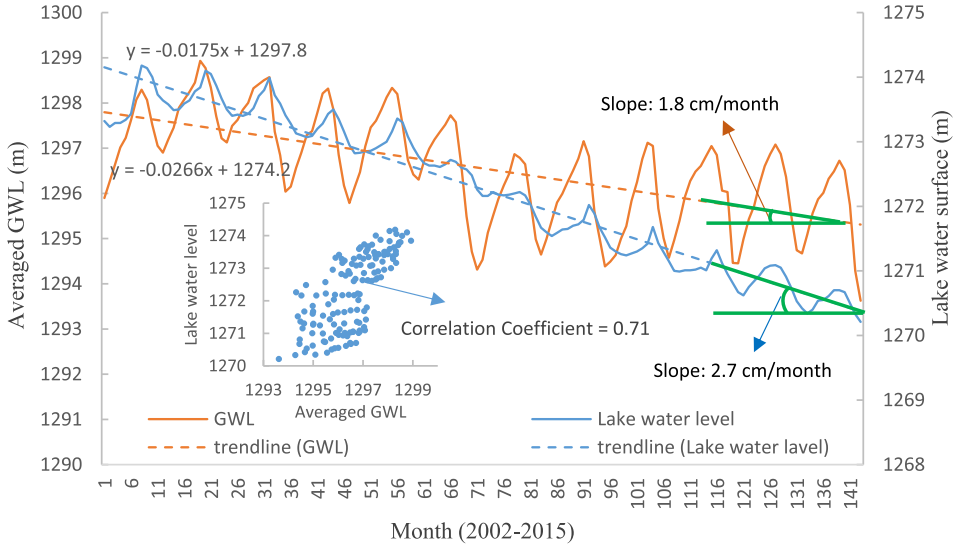

Fig. 1. Broad brush study of Lake Urmia water levels and the average GWL of all the Observation Wells (OW) provided by WAWA (2015).

model combination streams of activities are flanked by a traditional research practice, in which MMs are used but with the goal of comparing the performances of an innovative model with the performances of traditional MMs. The goal in traditional practices is to identify a superior model and this normally concludes with confirming the superiority of the innovative model by using anecdotal evidence and without statistical tests. The authors use various AIMM strategies and show significant improvements, e.g. see Nadiri et al. (2019).

The Bayesian paradigm gained popularity since the middle of the 20th century; whereas traditional model inter-comparison studies, available since the early 20th century, employ frequentist paradigms in one-way or another, see Jeffreys (1961). Newman and Wierenga (2003) introduced BMA to combine a number of conceptual groundwater models for flow and transport equations. Tsai and Li (2008) developed a new framework to address uncertainty associated with inverse groundwater modelling for hydraulic conductivity estimation. Further applications include: GWL prediction (Li and Tsai 2009) using a conceptual model; groundwater management under uncertainty to mitigate saltwater intrusion (Tsai 2010); Hierarchical BMA (HBMA) was applied to groundwater remediation designs to assess impacts of individual sources of uncertainty (Chitsazan and Tsai, 2014, 2015); also Chitsazan et al. (2015) quantified uncertainty in different ANN models by HBMA, mathematical models, and input data to predict fluoride concentration in groundwater.

Whilst BMA is a natural model combination strategy, it is fused in the paper with the model combination strategies based on the author's AIMM strategy for bottom-up data-driven techniques. Applications of Artificial Intelligence (AI) techniques without model combination are outlined in Table 1, which categorises various applications to groups. Research in Group 1 uses a single AI model but applies it to diverse case studies such as GWL prediction in karstic and leaky aquifers (Coppola et al., 2005; Trichakis et al., 2011); GWL prediction in freshwater swamp forests (Sun et al., 2016); and GWL prediction in an unconfined aquifer (Sadeghfam et al., 2018).

Research in Group 2 (Table 1) uses AI with MMs for GWL predictions using the same data, which often share the same goal of selecting a 'superior' model. As shown in Table 2, Bazartseren et al. (2003) use two such AI models and report that both models are accurate for a longer GWL prediction horizon. Alvisi et al., (2006) also use two AI models and report that Sugeno Fuzzy Logic (SFL) and Mamdani Fuzzy Logic (MFL) perform slightly better than that by Artificial Neural Network (ANN) model in terms of RMSE. Shiri et al., (2013) employ Gene

Table 1
List of literature review about GWL prediction by AI models.

	Researcher	AI Models tested		Key contributions
	Coulibaly et al., 2001	ANN		Trained 3 types of ANN to predict GWL fluctuations
Group 1: Use AI with Single models	Coppola et al., 2005	ANN		Use ANN to predict GWL in karstic and leaky aquifers Numerical models perform weakly in such cases
	Daliakopoulos et al., 2005	ANN		Trained 7 ANN architectures and algorithms
	Nadiri, 2007	ANN		Used ANN to evaluate FFN-LM over a complex aquifer
	Mohanty et al., 2010	ANN		Trained 3 ANN models for GWL predictions Predicted GWLs a tropical humid region
	Trichakis et al., 2011	ANN		Predicted GWL in a karstic aquifer
	Taormina et al., 2012	ANN		Predicted GWL in two-step for a coastal aquifer Use observed GWLs and external inputs.
	Chang et al., 2015	ANN		Trained ANN for Suprapermafrost GWL variation Studied response to climate change
	Sun et al., 2016	ANN		Trained ANN to predict GWL in freshwater swamp forest
	Sadeghfam et al., 2018	SFL		Trained Sugeno Fuzzy Logic (SFL) to predict GWL Filled in data gaps in GWL time series.
Group 2: Use AI with Multiple Models (MM) to Select the Best/Superior Model	Bazartseren et al., 2003	ANFIS	✓	Trained ANFIS and ANN to predict short-term GWLs
		ANN	✓	
	Alvisi et al., 2006	ANN		GWL predicted by FL and ANN.
		FL	✓	
	Szidarovszky et al., 2007	ANN		Introduced a hybrid of ANN and numerical GW flows
		GW flow model		Helped improving numerical model predictions
	Bisht et al., 2009	FL		Evaluated the efficiency of ANFIS and FL
		ANFIS	✓	Reported ANFIS to be superiority to FL models
	Jalalkamali and Jalalkamali, 2011	ANN		Trained NF and ANN
		ANFIS	✓	Combined monthly variabilities for GWL predicting.
	Moosavi et al., 2013	ANN		Tested predicting GWL with 1, 2, 3 and 4 months ahead
		ANFIS,		Tested various model structures
		Wavelet-ANN		Used ANN, ANFIS, Wavelet-ANN and Wavelet-ANFIS
		Wavelet-ANFIS	✓	Used wo test cases
	Shiri et al., 2013	GEP	✓	Used GEP, ANFIS, ANN and SVM techniques
		ANFIS		Evaluated GWL predicting GWL
		ANN		
		SVM		
	Emamgholizadeh et al., 2014	ANN		Used ANN and ANFIS models
		ANFIS	✓	Predicted GWL
	Tapoglou et al., 2014	ANN		Used ANN and Particle Swarm Optimisation (PSO)
		PSO		Evaluated GWL predictions under climate change scenarios
	Nie et al., 2016	ANN		Used RBF (Radial Basis Function) ANN and SVM
		SVM	✓	Predicted simulate GWL fluctuations.
	Gong et al., 2016	ANN		Used ANN, SVM and ANFIS
		SVM	✓	Predicting GWL interacting with GW
		ANFIS	✓	

✓ Denotes models claimed to be superior.
ANFIS: Adaptive Neuro Fuzzy Inference System.
ANN: Artificial Neural Network.
FL: Fuzzy Logic (SFL: Sugeno FL and MFL: Mamdani FL).
GEP: Genetic Expression Program.
GW: Groundwater.
PSO: Particle Swarm Optimisation.
SVM: Support Vector Machine.

Expression Programming (GEP) and report it to be more successful than other AI models used in their tests for GWL predictions.

The choice of models on groundwater problems is vast, and broadly comprise: (i) process-based theoretical/empirical modelling capabilities, such as MODFLOW, MT3D but are data-intensive and computationally expensive; or (ii) data-driven bottom-up modelling capabilities, such as AI techniques, which only use local data and still offer computationally cost-effective prediction capabilities. Notably, the combination of these two capabilities are referred to as conceptual models but are not detailed here for brevity. Owing to the sparsity of data and limited budget, AI techniques are natural choices for the study. Also, Nadiri et al. (2019) discuss the mathematical foundation for combining MMs, according to which model combination even by simple averaging improves the average of the combined model compared with the individual models. The techniques based on process-based modelling capabilities are precluded from this study but three commonly used AI models are selected to be combined by BMA to improve on their accuracy.

An overview of Table 1 provides a clear evidence that there is no superior model. Thus, such a selection is at the expense of discarding information contained in the remaining models often by attaching a lower status. The present paper is critical of choosing a superior or a single model at the expense of other models, as emphasised by Khatibi et al. (2011a, 2011b, 2013, 2014, 2017). Traditional model selection practices overlook the opportunity to learn from the convergence and divergences among MMs and they only provide anecdotal evidence for identifying the superiority of a model but in reality they are just fit-for-purpose. One approach to minimise the loss of information is to formalise the model selection procedure. For instance, Khatibi et al. (2003), Tilford et al. (2005) and Hawkes et al. (2005) discuss a risk-based approach to model selection; whereas, Li and Tsai (2009) discuss a Bayesian Model Selection (BMS) based on Bayesian model evidence

Table 2
Statistical features of data for each cluster.

Cluster		1	2	3	4	5	6	7	8	9	10	11	12
Mean	G	1277.9	1286.4	1301.6	1280.4	1293.2	1312.1	1278.9	1306.1	1348.3	1276.2	1287.1	1331.4
	T	11.61	11.61	11.61	11.61	11.61	11.61	11.61	11.61	11.61	11.61	11.61	11.61
	R	5.63	5.63	11.26	2.50	4.91	5.63	9.85	12.49	12.49	–	5.26	18.12
	A	0.24	0.61	0.04	0.12	0.40	0.00	0.37	0.76	0.29	0.38	0.85	0.31
	P	19.34	19.34	19.34	19.34	19.34	19.34	19.34	19.34	19.34	19.34	19.34	19.34
Variance	G	2.08	2.54	1.04	0.50	2.25	10.70	0.90	9.80	17.94	2.04	1.14	1.61
	T	76.34	76.34	76.34	76.34	76.34	76.34	76.34	76.34	76.34	76.34	76.34	76.34
	R	65.13	65.13	260.52	35.37	34.09	65.13	395.30	203.39	203.39	–	107.96	471.35
	A	0.13	0.80	0.00	0.05	0.33	0.00	0.29	1.21	0.18	0.28	1.50	0.20
	P	394.34	394.34	394.34	394.34	394.34	394.34	394.34	394.34	394.34	394.34	394.34	394.34
Maximum	G	1280.35	1288.64	1302.86	1281.96	1295.85	1315.39	1280.61	1310.78	1357.69	1278.7	1289.34	1333.72
	T	26.1	26.1	26.1	26.1	26.1	26.1	26.1	26.1	26.1	26.1	26.1	26.1
	R	43.27	43.27	86.54	41.63	38.62	43.27	121.04	95.52	95.52	–	63.38	138.79
	A	1.00	2.24	0.11	0.62	1.40	0.00	1.32	2.74	1.07	1.18	3.18	1.08
	P	88.6	88.6	88.6	88.6	88.6	88.6	88.6	88.6	88.6	88.6	88.6	88.6
Minimum	G	1274.58	1282.63	1298.38	1278.80	1290.01	1306.48	1276.79	1301.53	1339.91	1273.85	1284.57	1327.83
	T	−6.3	−6.3	−6.3	−6.3	−6.3	−6.3	−6.3	−6.3	−6.3	−6.3	−6.3	−6.3
	R	0	0	0	0	0.44	0	0	1.11	1.11	–	0	1.11
	A	0	0	0	0	0	0	0	0	0	0	0	0
	P	0	0	0	0	0	0	0	0	0	0	0	0

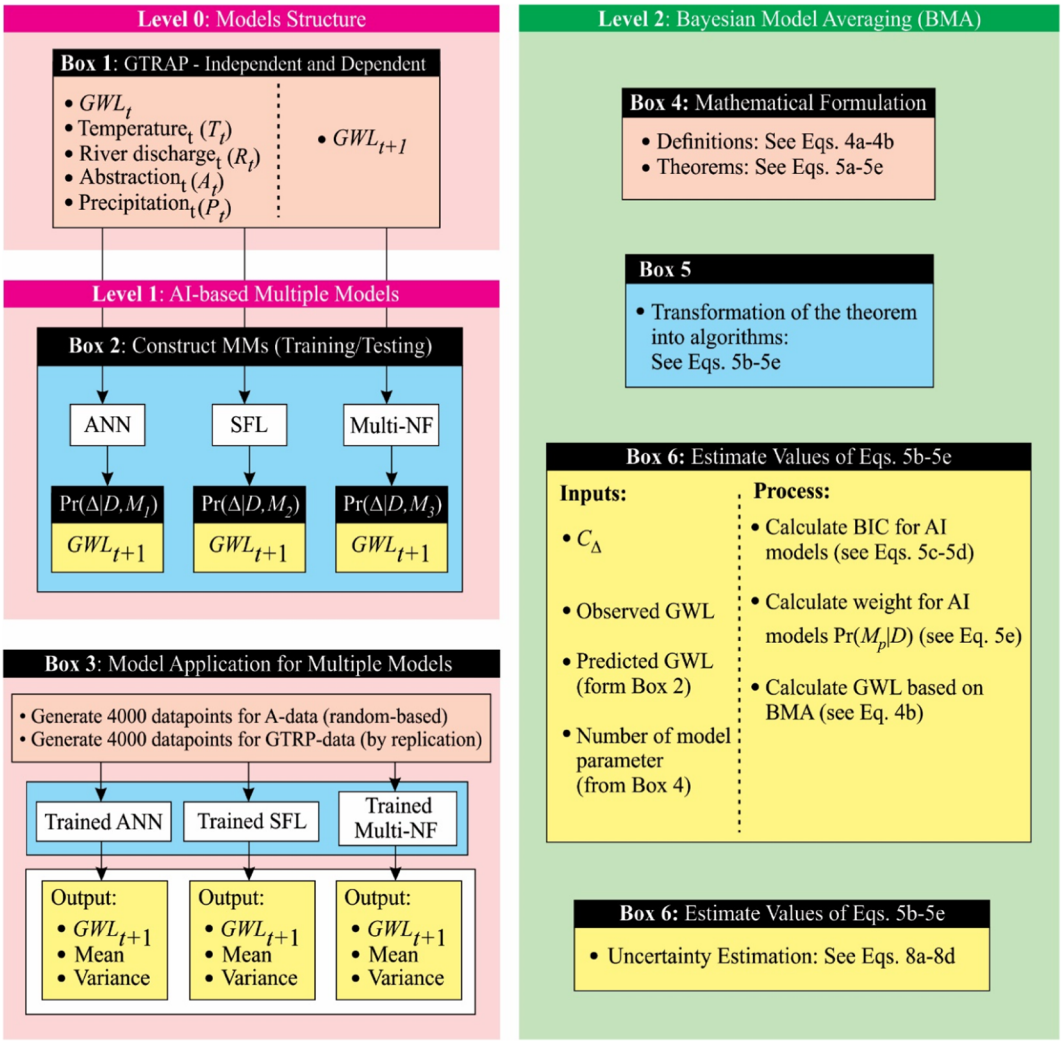

Fig. 2. Flowchart for the Methodology of BMA.

for selecting the best model and argue that the best model does not guarantee better prediction. Also, Tsai and Li (2008) argue that BMA provides robust results than BMS. Notably, both BMS and BMA use posterior probability to rank models (Link and Barker, 2006). Without testing the statistical significance of the superiority of a model, investigating BMA to combine AI models is a viable case for research.

The two streams of research reviewed above share the goal of model combination but those accounted by Group 2 in Table 1 require a strategy for model combination for a greater efficacy in modelling. The authors avoid ranking individual models for selecting a best/superior model; but promote learning from MMs through formulating a modelling strategy e.g. see Nadiri et al. (2014, 2018a, 2018b) and Khatibi et al. (2017, 2018a,b), Ghorbani et al. (2018). The capability to learn from MMs is implemented through formulating strategies, in which they are arranged at two levels: **at Level 1** a set of models are developed (which are Artificial Neural Network, Sugeno Fuzzy Logic, Neuro Fuzzy) by employing their best practice procedure but for the purpose of producing inputs to the next level; **at Level 2** a model is used to combine Level 1 models, the inputs of which are based on the outputs of Level 1 models. The model combination strategy used here is based on BMA. The only model combination strategy based on AIMM for GWL modelling is reported by Nadiri et al (2019). They formulate several strategies, including one by simple averaging of MMs and confirm the published literature by showing that the Root Mean Square Errors (RMSE) of the combined model is less than the mean of the RMSE of the individual models. This is a confirmation that even simple averaging improves accuracy.

The objective is to predict the water table (GWL) for the Urmia aquifer by combining AI models using BMA. The paper applyes BMA to combine different AI models for predicting GWLs and to the best knowledge of the authors, BMA combining MMs of AI has not been investigated for the prediction of GWL; also the model structure for prediction has novelty but this is explained in due course. The procedure involves intrinsically a capability for estimating uncertainty. Attention is drawn to a possible mismatch of terminology on 'observed input data' as used in Bayesian statistics, where these may be used as 'input data' in AI modelling practices. Both terms are used interchangeably in the paper.

2. Methodology

An overview of the methodology is outlined in Fig. 2 as a flowchart, which depicts the main elements of using BMA to learn from MMs for predicting Groundwater Levels (GWL). The main elements are described in this section, which lays out models at levels. As illustrated in Fig. 2, **Level 0** comprises decisions on model structures (Box 1). **Level 1** comprises the construction of MMs implemented in two phases of training and testing (Box 2); as well as their applications to set up a random-series to replace some of the observation data (Box 3). **Level 2** uses BMA to combine the MMs at Level 1 and to quantify inherent uncertainties (Box 6).

2.1. Level 0: Model structure

The approaches used to form the model structure include: (i) an autoregression expression of GWL in terms of its previous values; (ii) forming a set of regression equations; and (iii) their combinations. It suffices to state that GWL at $t + 1$ is defined as a function of observation data including GWL_t, Temperature (T_t), River discharge (R_t), groundwater Abstraction (A_t), Precipitation (P_t), see Box 1 of Fig. 2. These variables collectively form the acronym of GTRAP and the authors refer to it as a grey-box model, outlined in Section 3.3. A grey-box model of this nature is not common in the literature for groundwater problems. The authors are developing the approach, which is particularly suitable for cases with sparse data, see also Nadiri et al. (2019). These data were provided by West Azerbaijan Regional Water Authority (WAWA).

The GWLs over the basin are influenced by a diverse range of factors and are therefore GWLs are expected to be heterogeneous. Self-Organised Mapping is used to transform heterogeneous GWL records into a number of homogeneous datasets as clusters of Observation Wells (OW).

2.2. Level 1: Multiple models

The paper constructs the following MMs at Level 1 to predict GWL: (i) Artificial Neural Network (ANN); (ii) Sugeno Fuzzy Logic (SFL); (iii) Neuro Fuzzy (NF). As these models are now well-established, it suffices only to specify them to ensure that their implementation are reproducible by a third party. These models are implemented in two phases: training and testing.

2.2.1. Artificial Neural Network (ANN) – Box 2 in Fig. 2

ANNs are now well-established bottom-up data-driven modelling techniques with wide applications in hydrology. The ANN implementation in the paper is based on classic methodology as described by ASCE (2000a,b), Hornik et al. (1989), Haykin (1999), and Khatibi et al. (2017). Applications of ANN to the prediction of GWLs are cited in Table 1.

The implementation here comprises a feedforward Multi-Layer Perceptron (MLP) and uses three layers: (i) the input layer, which comprise 5 neurons of GWL_t (and more neurons on account of each OW at any given cluster) T_t, R_t, A_t, P_t (TRAP variables, which do not vary across the basin); (ii) the hidden layer, for which the number of neurons has to be determined through a trial-and-error procedure to be presented later; and (iii) the output layer, which comprises one neuron, GWL_{t+1}. Model building activities are processed in two phases: (i) the training phase using Levenberg-Marquardt (LM) algorithm, and (ii) the testing phase. These are expressed as (ASCE, 2000a,b):

$$O_j = f_1(b_j + \sum_i W_{ji} I_i) \tag{1a}$$

$$O_K = f_2(b_K + \sum_i W_{ji} O_j) \tag{1b}$$

where f_1 is the activation function of the hidden layer, and f_2 is its activation function of the output layer, I_i is the i^{th} input, O_j is the j^{th} output for hidden layer and O_k is the k^{th} output for output layer, W_{ji} and W_{kj} are weights that control the strength of connections between two layers, and the biases b_j and b_k are used to adjust the mean value for input layer and hidden layer, respectively. In this study, the Levenberg-Marquardt (LM) algorithm is used for the implementation of the backpropagation of the ANN model during its training phase.

2.2.2. Sugeno Fuzzy Logic (SFL) – Box 2 in Fig. 2

The Sugeno Fuzzy Logic (SFL) model (Takagi and Sugeno, 1985) is selected to form the second member of the MMs in this study and their applications for predicting GWLs are cited in Table 1. The first-generation FL-based modelling procedure was disadvantaged by relying on a set of prescribed rules; whereas SFL shifted the practice toward learning the rule-base from the site-specific data (machine learning) and overcame the subjectivity inherent in prescribed rule-bases.

SFL builds on ambiguous boundaries as in the classic fuzzy set theory. In contrast to crisp sets, gradual transitions are allowed between defined sets and partial membership function ranging from 0 to 1, for more details see Pulido-Calvo and Gutierrez-Estrada (2009) and Zadeh (1965). The advantage of fuzzy sets is that it copes with imprecision often inherent in hydrogeological parameters. The process of implementing the partial membership function involves using different shapes, such as triangular, Gaussian trapezoidal, sigmoid, S-shape and Z-shape, often selected through trial-and error at the preliminary stages.

SFL uses clustering techniques, which are available since the 1980 s.

Clustering learns automatically possible structures within the data and subsequently sets out optimum rules. It uses Subtractive Clustering (SC) by Bezdek (1981) and combines with FL, which comprises: (i) SC is used to extract fuzzy if-then rules (Chiu 1994; Chen and Wang 1999; Nadiri et al., 2013); and (ii) constant or linear output membership functions, referred to as the zero and first order SFL, respectively, are learned from the data. The clustering process is characterised by (i) the number of rules in SC, which is the same as the number of clusters; and (ii) the clustering radius, which is identified as a value between 0 and 1. Their final output is the weighted average of all rule outputs (aggregation), as follows (Bezdek, 1981):

$$GWL = \frac{\sum_{i=1}^{l} w_i GWL_i}{\sum_{i=1}^{l} w_i} \tag{2a}$$

where w_i = firing strength for the i^{th} rule, obtained through the "and" operator; and l = number of rules.

A first-order SFL model is implemented to model GWL values through the use of a generalised Gaussian function to express membership functions for the five-input data. Each input is clustered into appropriate classes and a set of fuzzy if-then rules are derived to linearly aggregate the input data as the output and these use the membership functions for the input data. For GWL at $t + 1$ predictions in this study, a fuzzy if-then rule i can be expressed as:

$$\text{Rule } i: \begin{Bmatrix} GWL \text{ belongs to } \mathrm{MF}^i_{GWL} \\ T \text{ belongs to } \mathrm{MF}^i_T \\ R \text{ belongs to } \mathrm{MF}^i_R \\ P \text{ belongs to } \mathrm{MF}^i_P \\ E \text{ belongs to } \mathrm{MF}^i_E \end{Bmatrix}, \text{ then } GWL_{i+1}$$
$$= m_i GWL + n_i T + p_i R + u_i A + q_i P + c_i \tag{2b}$$

whereGWL_{i+1} = output of rule i; $\mathrm{MF}^i_{\mathrm{GWL}}$ = membership function of the i^{th} cluster of input GWL; MF^i_T = membership function of the i^{th} cluster of input T; MF^i_R = membership function of the i^{th} cluster of input R; MF^i_A = membership function of the i^{th} cluster of input A; MF^i_P = membership function of the i^{th} cluster of input P and m_i, n_i, p_i, q_i, u_i and c_i = coefficients to be determined by the linear least-squares estimation.

2.2.3. Multiple Neuro-Fuzzy (Multi-NF) – Box 2 in Fig. 2

Neuro-fuzzy models integrate ANNs with fuzzy logic and therefore fuzzy inference rules are handled by ANN. Their applications to predicting GWLs are shown in Table 1. Thus, the integrated NF inference system comprise: (i) a given input dataset as specified above; (ii) an SFL whose MF parameters are tuned using a hybrid algorithm. The most compatible method for the construction of NF is the Sugeno method using subtractive clustering. The architecture of NF in this study comprises a five-layer MLP network as follows (Nadiri et al, 2013b):

Layer 1: Generate membership function for the input data, $X = \{GWL_t, T_t, R_t, A_t, P_t\}$, and the output of neuron i is expressed as:

$$O_i^{\ell=1} = \mu_{ji}(X) \tag{3a}$$

where $\ell = 1$ and indicate the Layer (it is not power, and this is also true up to Layer 5, as below), j is the number of inputs and i is the membership function index; $\mu_{ji}(X)$ is a fuzzy set associated with neuron i given a membership function. This study used a generalised Gaussian function for the MF function, which was identified through a trial-and-error procedure.

Layer 2: Calculate firing strength w_i for the i^{th} rule via the multiplication rule:

$$O_i^{\ell=2} = w_i = \mu_{1i}(X)\mu_{2i}(X)\mu_{3i}(X)\mu_{4i}(X)\mu_{5i}(X) \tag{3b}$$

Layer 3: Compute the normalised firing strengths for the i^{th} neuron:

$$O_i^{\ell=3} = \bar{w}_i = \frac{w_i}{\sum_i w_i} i = 1, \cdots, 5 \tag{3c}$$

Layer 4: Compute the contribution of the i^{th} rule in the model output using first-order SFL:

$$O_i^{\ell=4} = \bar{w}_i GWL_i = \bar{w}_i (m_i GWL + n_i T + p_i R + q_i A + u_i P + c_i) i = 1, \cdots, 5 \tag{3d}$$

Layer 5: Calculate the final output as the weighted average of all rule outputs (aggregation):

$$O_i^{\ell=5} = GWL = \sum_i \bar{w}_i GWL_i \tag{3e}$$

The NF parameters in Eq. (3d) and MF parameters are estimated using a hybrid algorithm in this study, which is a combination of the gradient descent and the least-squares method.

The implementation of NF in this research is referred to as a Multiple Neuro-Fuzzy (Multi-NF) model since the regular NF produce only one output. Multi-NF incorporates a number of NF arranged in parallel combination to achieve a model architecture with multiple output. This architecture is commonly applied in different fields (e.g. turbine runner analysis, see Saeed et al., 2013; optimal design of condenser, see Huang and Yu, 2016), but to the best of the authors' knowledge, this has not been applied to groundwater problems.

2.2.4. Implementations of models at Level 1

The input variables (GTRAP), defined in Section 2.1, and to be further detailed in Section 3.3, are each associated with uncertainty but the measurement of four of them (G-T-R-P variables) follow standard best practice procedures and their contribution to model uncertainty are unlikely to be dominant. However, groundwater abstraction (A) is a variable that contains significant uncertainty due to the scale of unaccounted abstractions. Monte Carlo Simulation (MCS) was used to generate 4000 random series for groundwater abstraction (Box 3 in Fig. 2). MCS employs uniform distribution in the range of the measured values of abstraction and a certain percentage from measured values to account for uncertainty stemming from unaccounted abstractions, the range between minimum and maximum unaccounted abstraction for each cluster are given later in Table 2 and the allowance is as per recommendation by WAWA (2015). These are sufficient to the application of the MMs to compensate for the insufficiency of data records.

2.3. Level 2: Bayesian model combination of multiple models – Boxes 4–7 in Fig. 2

2.3.1. Bayesian theorem and its transformation into an algorithm

Due to the complexity of the mathematical formulation given below, the presentation is first outlined in a simplified form but at the risk of a loss of rigour. The results of the three Level-1 models in a cluster may be represented as three arrays of $\{M_1\}$, $\{M_2\}$ and $\{M_3\}$ and their individual corresponding datapoints as (m_1), (m_2) and (m_3). Rather than using the Bayesian theorem to combine these models, consider their combination by simple averaging, e.g.:

$$(m_{combined}) = \frac{(m_1 + m_2 + m_3)}{3} = (0.33m_1 + 0.33m_2 + 0.33m_3)$$
$$= (w_1 m_1 + w_2 m_2 + w_3 m_3) \tag{4a}$$

The simple averaging technique assigns $w_1 = w_2 = w_3 = 0.333$ but instead of prescribed rules, the coefficients w_1, w_2, w_3 can be learned from the site-specific data, e.g. Nadiri et al. (2019). As introduced above, Bayesian statistics provide another strategy as investigated by the paper.

The Law of Total Probability: BMA combines n plausible models per cluster as expressed by Eq. (4b) below, which is a general form of

Eq. (4a) as follows, (Draper 1995):

$$\Pr(\Delta|D) = \sum_{p=1}^{n} \Pr(\Delta|D,M_P)\Pr(M_P|D) \tag{4b}$$

where $\Pr(\Delta|D)$ is the probability of the prediction of GWL (denoted as Δ) given the observed GWL (denoted as D); $\Pr(\Delta|D,M_P)$is the conditional probability of the predicted quantity given the observed data D and given model M_P; and $\Pr(M_P|D)$ is the posterior probability of the model, which are also known as model weight, given the data D, (see Draper 1995; Kass and Raftery, 1995; Raftery et al., 1997; Höting et al., 1999). In this study Δ represents predicted GWL, D denotes observed GWL and M_P denotes ANN, SFL and NF.

Bayesian Theorem: BMA uses the Bayesian theorem outlined through the following steps:

Step (i) – the theorem: The Bayes theorem, formulated for BMA, uses n plausible models $\{M_1, M_2, \ldots, M_n\}$, where each array is one representation of the state variable of predicting GWLs, and their corresponding observed GWL values at each of the observation well are denoted by D. The theorem is detailed by (Berger, 1985) and is expressed as:

$$\Pr(M_P|D) = \frac{\Pr(D|M_P)\Pr(M_P)}{\sum_{j=1}^{n}\Pr(D|M_j)\Pr(M_j)} \tag{5a}$$

where $\Pr(M_P|D)$ is the posterior probability, which learns a better estimate from the given data; $\Pr(M_P)$ is a prior model probability for the model M_p, evaluated by expert judgments or estimated e.g. Wöhling et al. (2015) and Elshall and Tsai (2014); $\Pr(D|M_P)$ is marginal likelihood function for model M_p.

Step (ii) – algorithms: Eq. (5a) is processed for each cluster as follows:

Prior Probability Pr(M_P): It represents a relative weight of one model against other models before visiting data (Ye et al., 2004). When there is not sufficient reason to prefer one model to another, Höting et al. (1999) and Wöhling et al. (2015) considered equal priority (reasonable, natural choice) to all models.

An Algorithm for Updating: Prior probabilities are updated to posterior probability given marginal likelihood function (also known as Bayesian model evidence). As per Li and Tsai (2009), marginal likelihood function is approximated by:

$$\Pr(D|M_P) \approx \exp[-\frac{1}{2}BIC_p] \tag{5b}$$

$$\text{where } BIC_p = Q_P + N\ln 2\pi + m_P \ln N \tag{5c}$$

$$\text{And in turn, } Q_P = (\Delta - D)C_\Delta^{-1}(\Delta - D)^T \tag{5d}$$

where N is number of Observation Well (OW) in a cluster; m_p is number of model parameters expressed by Eq. (7a) below; Q_p is the sum of weighted squared errors expressed by Eq. (5d); Δ is predicted GWL and D is measured GWL; C_Δ is the variance matrix of prediction errors using Monte Carlo simulations on model parameters (Li and Tsai 2009), see below.

Eqs. (5b)–(5d) are replaced in Eq. (5a) and are manipulated to derive the following:

$$\Pr(M_P|D) = \frac{\exp(-\frac{1}{2}\alpha\Delta\text{BIC}_p)}{\sum_{i=1}^{n}\exp(-\frac{1}{2}\alpha\Delta\text{BIC}_p)} \tag{5e}$$

where $\Delta\text{BIC}_p = \text{BIC}_p - \text{BIC}_{min}$; in which BIC_{min} (Bayesian Information Criteria) is the lowest BIC value among the models; α is a scaling factor used in the variance window. Using $\alpha = 1$ leads to Occam's window and this results in selecting the unique model as the best model; whereas, if $\alpha = 0$, the weights for identical models are obtained. Tsai and Li (2008) solved this problem by using a desired significance level in Occam's window as a scaling factor to produce the weights of a rational model. Reference may be made to Tsai and Li (2008) for selecting a value for the scaling factor.

Li and Tsai (2009) recommends BIC, as it gives an unbiased formulation and computational efficiency; it is able to penalise against over-parameterisation to ensure model parsimony; and it is widely applied. Published researches are indicative of investigating the performance of BIC, AIC (Akaike Information Criteria), and KIC (Kashyap Information Criterion), e.g. see Tsai and Li (2008), Singh et al. (2010), Tsai and Elshall (2013); although Liu et al. (2016) argue that their outcomes are often contradictory but this shortfall is overcome by using numerical methods such as Markov Chain Monte Carlo (MCMC) method. However, seeking a solution for this issue is outside the remit of the paper but it is understood that the choice of AIC suits complex models with large sample sizes over simple models.

Step (iii) – Monte Carlo Simulations: The Monte Carlo simulation performed by using the uniform probability distribution fitted to abstraction data. The paper makes an allowance for over-abstraction, as recommended by WAWA (according to which unauthorised abstractions vary up to 0.8 times of measured abstraction). Uniform probability distribution is used in preference to other types of PDFs due to insufficient data. Section 4.2 specifies the way for carrying out Monte Carlo simulation in this study but Eq. (6) involves C_Δ and its formulation is outlined as follows. As there are three models at Level 1, there are three variances, σ_i^2 per cluster, which estimates the variance between observed values and modelled values after the application of the Monte Carlo simulation. C_Δ per cluster is then expressed as (Li and Tsai 2009):

$$C_\Delta = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_i^2 \end{bmatrix} \tag{6}$$

where σ_i^2 is the i^{th} variance.

2.3.2. Estimating parsimony parameters

The number of AI model parameters is often large and adding more parameters increases the simulation accuracy but through overfitting. This is one obvious reason for model parsimony and is estimated by using the following equation (Nadiri et al., 2014; Heddam et al., 2012):

$$m_p = \begin{cases} (N_{input} + N_{output}+1)\times N_{hidden\ layer} + N_{output}, & \text{(ANN)} \\ (N_{input} + N_{output}) \times N_{rule} \times N_{MF}, & \text{(SFL)} \\ (N_{input} + N_{output}) \times N_{MF} + N_{input} \times N_{MF} \times 2, & \text{(Multi} - \text{NF)} \end{cases} \tag{7a}$$

where m_p is number of model parameter, N_{input}, N_{output} are the number of input and output parameters, and $N_{hidden\ layer}$ is number of neurons in the hidden layer. N_{MF} represents the number of membership function for each input. Notably, the numbers vary from one cluster to another.

Ye et al., (2004) use a parsimony parameter, λ and is used in BMA, which is expressed as:

$$\lambda = m_p \ln N \tag{7b}$$

where N is the number of OW in each cluster. The number of each of the parameters, λ, Q_p, BIC and posterior quantities may seem to three as "1" per each model at Level 1 but this not true. The actual number accounts for the number of clusters and this is discussed further in Section 4.1.1.

2.3.3. Estimating uncertainties using BMA

The Bayesian model combines AI models of Level 2 using three measures: (i) the model parsimony as expressed by Eq. (7b); (ii) uncertainties associated with measured abstraction and referred to as *within model uncertainty*; and (iii) *uncertainties between AI models*. According to the law of total expectation and variance, the mean and

variance of the predictions are respectively (Draper, 1995):

$$E(\Delta|D) = \sum_{p=1}^{n} E(\Delta|D,M_p)Pr(M_p|D) \quad (8a)$$

$$Var(\Delta|D) = E_M Var(\Delta|D,M_p) + Var_M E(\Delta|D,M_p) \quad (8b)$$

where Eq. (8a) comprises two terms and its first term expresses the within-model variance as:

$$E_M Var(\Delta|D,M_p) = \sum_{p=1}^{n} Var(\Delta|D,M_p)Pr(M_p|D) \quad (8c)$$

Also, the second term expresses the between-model variance as:

$$Var_M E(\Delta|D, M_p) = \sum_{p=1}^{n} [E(\Delta|D, M_p) - E(\Delta|D)]^2 Pr(M_p|D) \quad (8d)$$

3. Study area and data

3.1. The aquifer system

Urmia plain, West Azerbaijan province, stretches along the west coast of Lake Urmia, northwest Iran (Fig. 3). It covers an area approx. 1000 km^2 and overlaps with the aquifer, which encompasses the basins of the following main rivers: *Shaharchay*, *Nazluchay*, *Rozechay*, and *Baranduzchay*. The rivers rise approx. along the borders between Turkey and Iran at the mountain ranges known as *Mor Daghlari* with notable mountains of *Sero*, *Silvana*, *Movana* and *Ziveh*. The watercourses flow in the general easterly direction towards Lake Urmia at the east through Urmia plain. Currently both the rivers and the aquifer suffer encroachments onto their natural states due to the absence of a planning system. This has resulted in serious impacts including serious declines in GWLs, some aspects of which are reported by Amirataee and Zeinalzadeh (2016). *Shaharchay* passes through the historic city of Urmia, the provincial capital of West Azerbaijan, at the centre-west of the plain. Average altitude of the plain is 1320 AMSL.

Urmia aquifer is the source for 26 qanats, 15 springs and 18,745 pumps, where the installation of the latter components goes to the years since 1990. Estimates indicate that the extractions were as much as 585 million m^3 of water from the aquifer in 2004–2005. The study area is characterised as a semi-arid and cold climate based on Emberger method (1930). The average annual precipitation at Urmia station is 350 mm and its mean annual temperature is 12 °C for a 10-year period (2004–2014) record (provided by WAWA). As per Iranian Meteorological Organisation, the mean annual relative humidity at Urmia city is 58% (2004–2014).

The aquifer at Urmia plain is a single formation layer, even though it is crossed by four significant watercourses (a report by the WAWA (2015)). The aquifer is unconfined and comprises heterogeneous alluvial deposits (see Fig. 4a) with high specific yields and a known high quality groundwater, which is one of the 12 aquifers around Lake Urmia. The location of 48 OWs on the young terraces and gravel plain is shown in Fig. 4a. Notably, these OWs were classified into clusters to treat their heterogeneity, using a range of parameters and the study identified 12 clusters by using Genetic Algorithm – Self Organising Map (GA-SOM) method. SOM is now widely used in hydrology e.g., Hsu et al. (2002) and Kalteh et al. (2008). SOM study of the study area is being reported separately but the clusters are presented in Fig. 4b.

3.2. Data availability

GWL monitoring is relatively new and is the responsibility of West

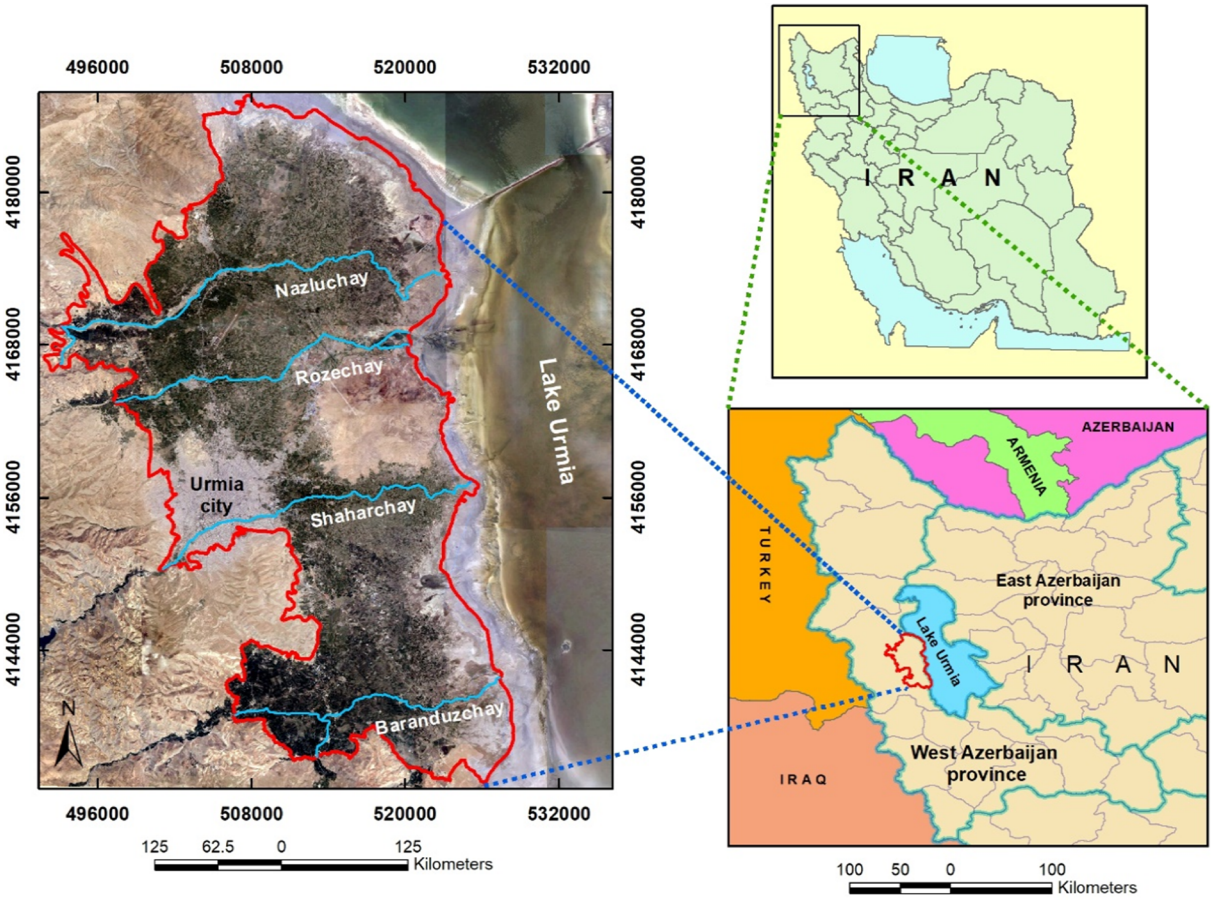

Fig. 3. Location map of study area.

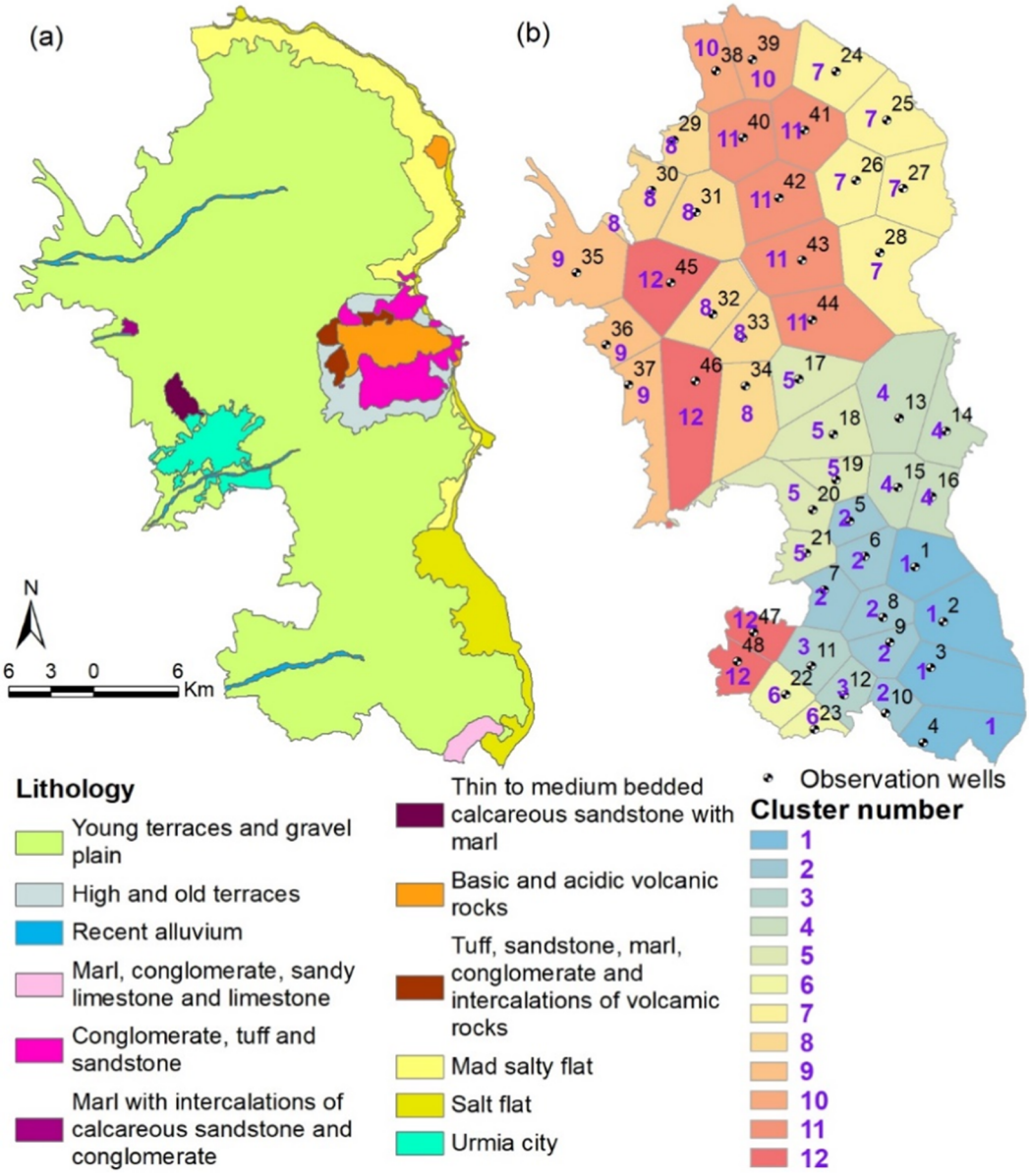

Fig. 4. (a) Lithological map of study area; (b) spatial location of clusters and observation wells.

Azerbaijan Water Authority (WAWA). The records go back to 2002 and these were obtained in 2015 for carrying out this research works. The monitoring network constitutes 71 OWs, of which 48 are active now with monthly readings of GWL and the available data for the study cover the period of 2002–2015. Further data available are the monthly precipitation and daily temperature data, both for the period 1967–2015 at the Meteorological Station in Urmia. Finally, daily river discharge data are available for the period of 1967–2015 from 8 stations in the 4 rivers. Table 2 presents basic statistics of GTRAP for each cluster including their mean and variance.

3.3. Model structure and the dataset

A grey box model is developed to prediction GWL in terms of a set of widely available characteristics, the input for which are GTRAP variables of: (i) one time-step lag of GWL time series (GWL_{t+1} expressed in terms of GWL_t), (ii) Temperature (T_t), (iii) River discharge (R_t), (iv) groundwater Abstraction (A_t) and (v) Precipitation (P_t). These observation data were used to form the dataset for OWs in parallel for each cluster and their values for each variable were normalised between 0 and 1. This ensures that one model is fit for each cluster.

The datasets were prepared by dividing randomly the data within each cluster into two parts as follows, 80% was used for the training phase and 20% for the testing phase. The measured values of the GTRAP independent variables are used in the training and testing phases of the Level 1 models. However, when observation data are prepared for feeding into the Level 2 Bayesian model, the variable A (Abstraction) uses its randomised-series, generated by the uniform distribution method. The assumptions are based on the advice given by WAWA, according to which the maximum is set to 180% and the minimum to 0% of measured values.

3.4. Performance measures

AI model training/testing and results of the four models use the following performance measures: (i) the Root Mean Square Error (RMSE) criterion, which has the value of 0 for a perfectly-fitted model and has no upper limit, although the lower the value, the better is the performance; and (ii) determination coefficient (R^2) with a best value of 1 for the perfectly-fitted model but when its value is zero, its

Table 3
Goodness of fits of AI models for 48 OWs cluster.

Cluster No.	O. W. No.	Max A (m^3/s)	Min A (m^3/s)	ANN		SFL		NF		BMA	
				R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Cluster 1	1	0.80	0.08	0.899	0.103	0.962	0.065	0.902	0.102	0.966	0.062
	2			0.908	0.098	0.959	0.067	0.934	0.084	0.962	0.065
	3			0.918	0.101	0.962	0.068	0.922	0.098	0.964	0.067
	4			0.947	0.069	0.961	0.060	0.941	0.073	0.975	0.048
Cluster 2	5	1.79	0.26	0.921	0.078	0.984	0.035	0.949	0.067	0.984	0.037
	6			0.935	0.086	0.978	0.051	0.960	0.068	0.991	0.035
	7			0.936	0.086	0.987	0.038	0.937	0.085	0.990	0.037
	8			0.959	0.085	0.984	0.053	0.980	0.061	0.993	0.037
	9			0.963	0.081	0.986	0.050	0.976	0.066	0.995	0.031
	10			0.956	0.080	0.980	0.055	0.960	0.076	0.982	0.052
Cluster 3	11	0.09	0.02	0.903	0.113	0.931	0.095	0.935	0.093	0.940	0.091
	12			0.876	0.097	0.918	0.078	0.903	0.085	0.908	0.083
Cluster 4	13	0.49	0.01	0.857	0.093	0.950	0.056	0.906	0.078	0.907	0.077
	14			0.900	0.076	0.961	0.048	0.911	0.075	0.926	0.067
	15			0.942	0.074	0.989	0.032	0.963	0.059	0.966	0.057
	16			0.933	0.084	0.981	0.045	0.944	0.078	0.961	0.066
Cluster 5	17	1.12	0.19	0.875	0.103	0.969	0.052	0.936	0.075	0.966	0.056
	18			0.961	0.060	0.985	0.037	0.976	0.047	0.992	0.029
	19			0.958	0.071	0.980	0.049	0.969	0.062	0.994	0.029
	20			0.972	0.054	0.980	0.045	0.982	0.044	0.992	0.030
	21			0.952	0.076	0.983	0.045	0.908	0.115	0.961	0.071
Cluster 6	22	0.002	0.0004	0.917	0.083	0.910	0.086	0.885	0.099	0.930	0.076
	23			0.956	0.087	0.948	0.093	0.941	0.100	0.963	0.079
Cluster 7	24	1.06	0.20	0.913	0.084	0.958	0.060	0.930	0.080	0.982	0.041
	25			0.863	0.106	0.942	0.070	0.870	0.104	0.980	0.045
	26			0.916	0.101	0.965	0.066	0.959	0.072	0.988	0.041
	27			0.861	0.124	0.958	0.069	0.910	0.102	0.973	0.058
	28			0.915	0.096	0.974	0.054	0.963	0.064	0.990	0.035
Cluster 8	29	2.19	0.38	0.909	0.094	0.971	0.054	0.940	0.077	0.979	0.046
	30			0.797	0.106	0.974	0.026	0.853	0.077	0.579	0.170
	31			0.951	0.068	0.978	0.046	0.970	0.054	0.992	0.033
	32			0.937	0.089	0.991	0.034	0.984	0.045	0.996	0.030
	33			0.859	0.130	0.970	0.061	0.901	0.111	0.987	0.043
	34			0.949	0.085	0.987	0.042	0.964	0.071	0.996	0.024
Cluster 9	35	0.86	0.12	0.945	0.076	0.966	0.056	0.955	0.066	0.971	0.054
	36			0.930	0.071	0.969	0.047	0.971	0.045	0.975	0.043
	37			0.955	0.056	0.970	0.045	0.970	0.045	0.973	0.043
Cluster 10	38	0.94	0.21	0.981	0.059	0.986	0.052	0.956	0.093	0.993	0.038
	39			0.966	0.063	0.957	0.074	0.969	0.059	0.984	0.043
Cluster 11	40	2.55	0.44	0.931	0.080	0.986	0.036	0.975	0.048	0.986	0.037
	41			0.918	0.074	0.965	0.050	0.955	0.056	0.982	0.038
	42			0.947	0.065	0.982	0.037	0.944	0.067	0.990	0.029
	43			0.948	0.078	0.984	0.044	0.948	0.079	0.981	0.048
	44			0.959	0.056	0.975	0.044	0.954	0.062	0.990	0.028
Cluster 12	45	0.87	0.15	0.896	0.113	0.912	0.106	0.883	0.124	0.950	0.080
	46			0.881	0.103	0.903	0.101	0.870	0.109	0.934	0.080
	47			0.863	0.093	0.826	0.122	0.862	0.097	0.950	0.059
	48			0.878	0.094	0.934	0.072	0.859	0.104	0.950	0.062
Colour Code						Good		Poor			

performance is poor or indicates no correlation.

4. Result

This section presents the results for both the models at Level 1 (ANN, SFL and Multi-NF) and the outcomes of the Bayesian model at Level 2. The section also specifies the decisions made at the stage of preliminary models, at the phase of training/testing and at their application. Attention is drawn to the implication of the clustering technique, as a result of which each of the four models (ANN, SFL, Multi-NF, and the Bayesian model) are processed as per cluster arranged in parallel and the dataset within each of the 12 clusters are also in parallel.

4.1. Prediction using AI multiple models at level 1

4.1.1. Artificial Neural Network (ANN)

A trial-and-error procedure was used to select the number of neurons in the hidden layer and the preliminary tests indicated that the architecture of 4 neurons at the hidden layer of all MLP-LS models would produce optimum performance metrics. The number of neurons for input and output layers depend on number of OWs in each cluster. Consider the example of Cluster 1, which contains 4 OWs and therefore its architecture is 8-4-4 as follows: neurons in the input layer: 8 (G_{OW1}-G_{OW2}-G_{OW3}-G_{OW4}-T-R-A-P); neurons in the hidden layer: 4; and neurons in the output layer: 4 (GWL_{t+1} for OW1, OW2, OW3 and OW4). By the same token, the number of input neurons for each cluster depends on the number of its OWs. Table 3 presents performance metrics for each OW and Fig. 5 exemplifies monthly-GWL predictions by ANN at 6 selected OWs for the 14 years of recorded period. The full results, presented in Supplementary Electronic Material, provide evidence that ANN performances are fit-for-purpose.

4.1.2. Sugeno Fuzzy Logic (SFL)

Using the Subtractive Clustering (SC) method, the values of cluster radius and the number of rules were identified by minimising RMSE between observed and predicted GWLs and these values are presented in Table 4 for each cluster. The Gaussian membership function was used to fit the data using their mean and standard deviation, although other functions were also tested. Table 3 presents the performance metrics for each OW and Fig. 5 exemplifies the monthly-GWL predictions by SFL at the 6 selected OWs against their observed GWLs for 14 years of the recorded period. The full set of the results are presented in Supplementary Electronic Material. They provide evidence that the performance of SFL is fit-for-purpose.

Table 4
SFL parameters: radius clustering and number of rules for each cluster.

Cluster	Number of OW	Optimal radius	Number of rules
1	4	0.4	49
2	6	0.4	92
3	2	0.8	5
4	4	0.4	76
5	5	0.4	94
6	2	0.8	4
7	5	0.4	103
8	6	0.5	78
9	3	0.9	5
10	2	0.8	6
11	5	0.4	93
12	4	0.9	8

4.1.3. Multiple Neuro Fuzzy (Multi-NF)

The NF model incorporated the same clusters through SC method for input and target datasets as those for SFL. The model parameters were estimated through the combination of the backpropagation gradient descent method (Jang et al., 1997) with the least-squares method (see Eq. (3d)) and the Gaussian membership functions. The NF implementation for the study takes account of multiple OWs in each cluster in parallel. Notably, the same model structure is used for all the OWs within the same cluster but the inputs and outputs are differentiated as per OW and not as per specific cluster, hence and hence Multi-NF. Table 3 presents the performance metrics for each OW located in each cluster and Fig. 5 exemplifies monthly-GWL predictions by Multi-NF for the 6 selected OW for the 14 years of the recorded period. The full set of the results are presented in Supplementary Electronic Material. They provide evidence that the performance of Multi-NF is fit-for-purpose.

Table 3 identifies the most representative (best performance) and the poorest performing models, according to which there is no single model to perform best in all cases and arguably this is a norm that its truth by-and-large will emerge sooner or later in the modelling culture to defy the notion of superior models. Overall, SFL performs better than NF and it performs better than ANN but not as good as SFL. Generally, all three sets of results are fit-for-purpose.

4.2. Predictions at level 2 and within-model uncertainty using MMs at level 1

Monte Carlo simulations were used in the model application process to generate the predictions by using random series of observation

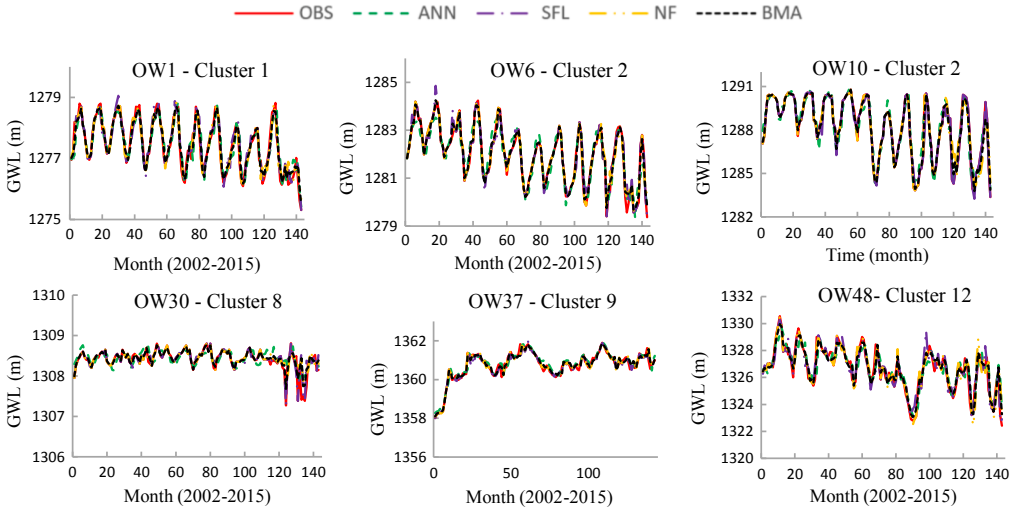

Fig. 5. Predicted and observed GWL by AI models and BMA.

dataset to capture the within-model uncertainty. However, randomisation was not applied to GTRP data not only to the A (Abstraction) data using a randomised-series of 4000 datapoints to produce the outputs from the MMs at Level 1. The 14 years of data records for each GTRP variable at each OW comprises only 168 datapoints; whereas predicted Abstractions comprise 4000 datapoints. This mismatch is dealt with by replicating GTRP-variables sequentially to make 4000 datapoints. These were fed into the Bayesian model at Level 2 but without showing prediction results of MMs at Level 1.

The Monte Carlo simulation of the values of the abstraction variable, using the uniform probability distribution, required a number of decisions, as follows: (i) the minimum values of over-abstraction was set to '0' and (ii) the maximum value was set to 80% greater than the measured value of abstraction, as recommended by WAWA (2004). Owing to the lack of any reliable information on the distribution of this variable, sensitivity tests of using other distributions are not justifiable.

4.3. Results of the Bayesian model

The Bayesian model is implemented in 3 steps as illustrated in Boxes 4–6 of Fig. 2 and their results are elaborated in this section. The BMA model is implemented as follows: (i) generate the prediction results from the Level 1 MMs and using Monte Carlo simulations for the Abstraction component (Box 3, Fig. 2), (ii) implement BMA by processing Eqs. (5a)–(5e) and (6) for C_Δ (a $n \times n$ diagonal matrix for a cluster, where n is number of OWs located in a cluster); (iii) produce the Bayesian predictions per cluster but capable of producing the results for each well, as per Eq. (4b). BMA learns from site-specific data, which is expected to reflect the observed values with more accuracy. The results presented in Tables 3, 5 and 6 and Fig. 5 exemplify prediction time series at 6 selected OWs. Evidently, the results are fit for-purpose.

The Bayesian model predictions include the three quantities of: λ as expressed by Eq. (7b); Q_P, as expressed by Eq. (5d) and measures the joint effect of input parameter uncertainty and between-model uncertainty; and BIC_p, as expressed by Eq. (5c) for each cluster of OWs (see also Fig. 2, Box 6). Notably, it is known that in existing implementations of the Bayesian models, λ becomes a dominant factor over BIC_p and Q_P (Nadiri et al. 2014), and this reduces the contributions of these latter two factors. The paper treats this known problem by normalising the number of parameters in each of the MMs between 0 and 1 and the production of the actual prediction results in the paper uses this novel treatment to smoothen model runs.

The values of the above three parameters of λ, Q_p and BIC together with the weights are presented in Table 5. These in turn are used to produce predicted values of the performance metrics are presented in Table 6, according to which the Bayesian model performs better than each of the MMs at Level 1 for 32 OWs out of 42 but the Level 1 models perform better in the case of the remaining 10 OW. Notably, the paper does not encourage traditional rankings.

4.4. Inter-comparison of the results

Without any encouragement to rank the performance of the four MMs (ANN, SFL, NF and Bayesian predictions), some insight is gained into the result by presenting the performance metrics for each cluster of the four models in Table 6. It also shows the values of the weights learned for each model from the measured data, as well as the values of the weights for each model at Level 1. Notably, Table 6 is a summary of Table 3, which the latter presents the performance metrics for each OW. The results suggest that there is some discordance between the values of the posterior probabilities and performance metrics of the individual models at Level 1. For instance, SFL performs better than ANN and NF in terms of performance metrics, but these do not translate into higher posterior probabilities for SFL. This could be attributed to: (i) not using more sophisticated probability distribution function for the Monte Carlo simulation due to the inherent level of uncertainties, but more sophisticated distribution functions are not justifiable; (ii) a general expectation is that locally learned values are often locally optimum values and often are not transferrable to meaningful inferences. The overall improvements are deemed to be good enough but further updating is justifiable with more data.

Time Plots: The above partial view on inter-comparison of the MMs are based on performance metrics but a further insight emerges when GWL hydrographs are displayed, as in Fig. 6. The obvious observation is that despite some of very good metrics, the deviation between observed and modelled results are significant. Fig. 6 shows that an AI model with the highest performance criteria is not the superior model in all time periods, and an AI model with the lowest performance criteria is far too good to be rejected. For example, consider OW 48 at Cluster 12, at which SFL is the best model in terms of performance metrics but as it is visible from the figure, it performs badly with respect to observed GWL between 95th and 100th months. The full set of the results are presented in Supplementary Electronic Material.

Scatters in the Results: The scale of deviations is more visible in the scatter diagram given in Fig. 7. The figure exemplifies another reality by showing the scatter diagram of 6 selected OWs that the model with highest performance criteria does not provide the best result in all time period and the model with lowest criteria are often fit-for-purpose.

A greater visual insight emerges by considering the scatter diagram of the error residuals defined as observed values minus prediction values. A sample of 6 OWs are shown in Fig. 8 but the full set of the results are presented in Supplementary Electronic Material. These results

Table 5
BIC and related terms and the calculated weights for AI models of different clusters.

Cluster no.		1	2	3	4	5	6	7	8	9	10	11	12
λ	ANN	0.07	0.02	0.10	0.072	0.02	0.12	0.02	0.02	0.19	0.10	0.02	0.19
	SFL	1.16	1.29	0.33	1.18	1.08	0.32	1.16	1.34	0.48	0.38	1.14	1.38
	Multi-NF	0.15	0.48	0.26	0.13	0.51	0.25	0.43	0.43	0.46	0.20	0.45	0.68
Q_p	ANN	0.31	4.10	0.64	15.66	3.14	0.75	3.56	5.43	2.13	0.91	3.26	1.51
	SFL	2.34	2.78	4.57	32.09	2.44	1.19	2.59	4.16	2.17	1.37	2.82	3.43
	Multi-NF	0.40	1.60	0.71	16.27	1.61	1.20	1.35	2.49	1.69	0.81	1.78	2.02
BIC	ANN	7.59	15.11	4.21	22.94	12.31	4.30	12.73	16.44	7.46	4.48	12.42	8.66
	SFL	8.53	12.52	7.91	38.26	10.55	4.55	10.62	13.85	7.20	4.66	10.87	9.40
	Multi-NF	7.60	12.14	4.12	23.49	10.29	4.62	10.10	13.09	6.74	4.28	10.52	8.68
Posterior Probabilities	ANN	0.39	0.13	0.47	0.57	0.17	0.38	0.14	0.12	0.27	0.33	0.18	0.38
	SFL	0.23	0.40	0.03	0	0.39	0.32	0.38	0.37	0.31	0.29	0.38	0.25
	Multi-NF	0.38	0.47	0.50	0.43	0.44	0.30	0.48	0.51	0.42	0.38	0.44	0.37

Note 1: Model weights are based on α in the BIC values and the paper uses $\alpha = 2.12/\sqrt{n}$ (n = number of OW, $2\sigma_D$ variance window size and 5% significance level).
Note 2: The number of each of the parameters, λ, Q_p, BIC and posterior quantities, are 12 for each of the models (ANN, SFL and Multi-NF) at Level 1 due to clustering. Therefore, the number of each of these parameters is a total of 36.

Table 6
Correspondence between performance metrics of Level 1 models and their posterior model probabilities of AI models for each cluster.

Cluster no.	ANN		SFL		NF		Posterior Probabilities			BMA	
	RMSE	R2	RMSE	R2	RMSE	R2	ANN	SFL	NF	RMSE	R2
Cluster 1	0.107	0.920	0.065	0.962	0.132	0.927	0.385	0.233	0.382	0.062	0.967
Cluster 2	0.107	0.952	0.047	0.984	0.122	0.965	0.130	0.399	0.471	0.038	0.990
Cluster 3	0.105	0.894	0.087	0.929	0.113	0.926	0.470	0.030	0.500	0.080	0.940
Cluster 4	0.089	0.920	0.046	0.975	0.096	0.939	0.573	0	0.427	0.076	0.932
Cluster 5	0.093	0.949	0.046	0.980	0.122	0.951	0.170	0.390	0.440	0.048	0.979
Cluster 6	0.105	0.943	0.089	0.936	0.102	0.920	0.383	0.317	0.300	0.077	0.952
Cluster 7	0.126	0.900	0.064	0.962	0.118	0.932	0.139	0.378	0.483	0.045	0.983
Cluster 8	0.114	0.913	0.087	0.981	0.120	0.952	0.120	0.369	0.511	0.077	0.970
Cluster 9	0.077	0.956	0.050	0.976	0.066	0.973	0.269	0.315	0.416	0.047	0.979
Cluster 10	0.057	0.975	0.064	0.974	0.062	0.979	0.329	0.289	0.382	0.041	0.989
Cluster 11	0.091	0942	0.042	0.959	0.088	0.955	0.179	0.377	0.444	0.036	0.985
Cluster 12	0.098	0.897	0.102	0.906	0.143	0.887	0.375	0.254	0.371	0.071	0.952
Colour Code		Orange: High Contribution			Green: High Contribution				Blue: Best performance		

clearly show the behaviour of each set of the models at each observation well. An examination of the full set of results show that the scatters are (i) narrow at some 24 OWs; (ii) medium at some 13 OWs; (iii) large at other 13 OWs; and (iv) there 5 OWs, where the scatters seem cluster or complex. As the length of time for the recorded data is 14 years, this is considered as sufficiently short and not capable of studying them for any pattern on trends.

Highlights: The highlights of the overall results presented in this sections are as follows: (i) no single model performs the best in all cases and this confirms Table 1, despite the rhetoric of claims on superior models; (ii) performance metrics are useful summaries and together with scatter diagrams they uncover the aspects hidden by performance metrics that the fitted models are hardly perfect; (iii) much clearer insight is obtained by scatter diagram of error residuals and the results show that BMA is effective in making the scatters narrower than those of the other models and thereby more robost, also there are patterns of behaviour; (iv) overall, narrow and medium residuals generally correspond to areas where the decline in water table is considerable (OW32, OW34 and OW38 for the narrow scatter, OW35, OW31, OW24 for the medium scatter), although some OWs with this behaviour have mixed scatter (see OW36 and OW23); (v) in the area with weak declines in water table, scatters are seen to be narrow, medium, large or in cluster forms.

5. Discussion

The results presented above clearly show that each set of performance measures reveal some aspects of the problem and therefore one is not a replacement for another but they have to be used side-by-side for a deeper insight. For instance, relying on performance metrics of

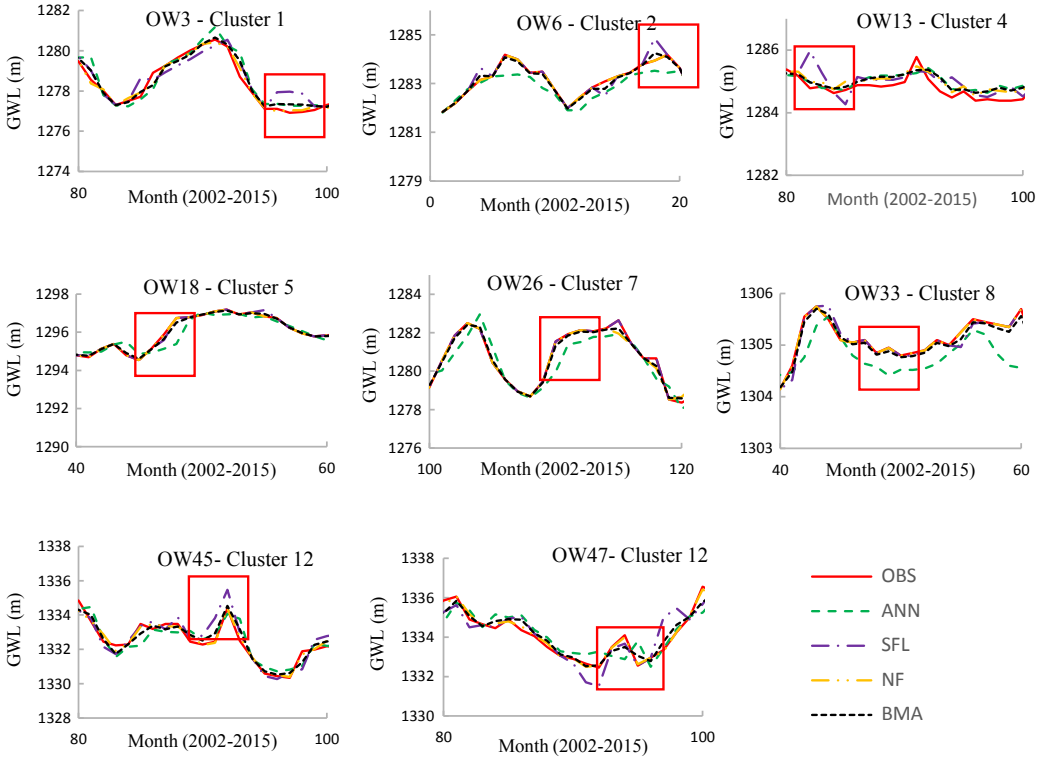

Fig. 6. Further views of the predicted and observed GWL time series.

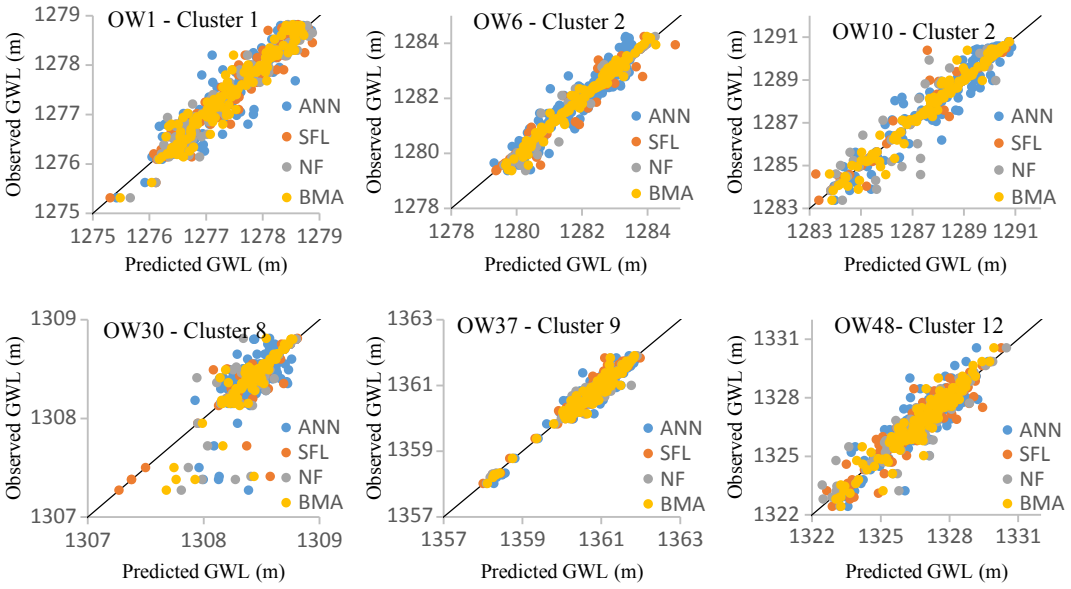

Fig. 7. Scatter diagram using observed and prediction GWL by AI models and BMA.

RMSE and R^2 reveal that each of the four models investigated in this study are fit-for-purpose but BMA tends to provide narrower error residuals than any other model. The strategy based on BMA is a two-level model management strategy, the various practices for which are reviewed in the Introduction section and is evidently unlike existing practices of selecting one at the expense of the others. The whole procedure gives rise to hierarchically arranged models (Levels 1 and 2) and renders: (i) it improves modelling accuracy and reliability; and (ii) a learning environment, and as such, a modelling strategy is not formed for the sake of finding a best model but for learning. Models are tools of decision making and learning in the following ways.

Theoretical basis for the improvements: There are theoretical reasons for improved performances of the model at Level 2 compared with the performances of MMs at Level 1. The reasons are based on Cauchy's inequality, as presented by Chen and Lin (2006), Kadkhodaie-Ilkhchi et al. (2009) and Nadiri et al. (2018a). However, their mathematical treatments are outside the scope of the paper.

Spatial distribution of the improvements: BMA is capable of quantifying the *within-model variance* owing to producing different predictions as per each model in MMs defined by Eq. (8c) and *between-model variance* in terms of assessing the convergence/divergence between the results of different models using Eq. (8d), both for the same study. The spatial distribution of both variances are presented in Fig. 9a-c, which uses the Ordinary Kriging (OK) method to distribute these values. The figure shows that: (i) in vast areas of the aquifer under study the data do not show any significant variance; (ii) modelling results are convergent in terms of having low variances associated with between-model variances; (iii) central areas of the aquifer under study are prone to both high within-model and between-model variances and this indicates there are significant divergences in these areas; and (iv) the strong similarity in the distribution of both components of variance is reflected in their total values; (v) the similarity between the total variance and the decline of water table shown in Fig. 9d is also striking. Notably, Fig. 9d is the spatial distribution of the results presented in Fig. 1 and also uses OK to distribute spatially the average declines. Land use in the central west of the aquifer under study comprises fruit farms with high demands on water supply.

A further attention is drawn to Fig. 9 to apparent similarities between Fig. 9a and 9b. Past experience on applications of BMA to groundwater models are largely based on predicting parameters, which derive a single value at sparse spatial locations distributed often by OK. Using past experience, one therefore may expect that considerable differences between the results in Fig. 9a and 9b. However, the task in this study is quite different, as BMA is used to predict time series, which

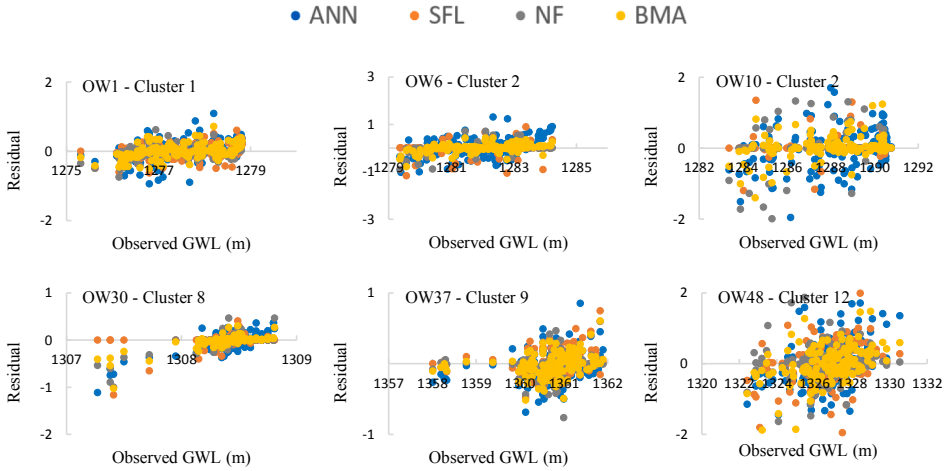

Fig. 8. Scatter Diagram of error residuals at 6 selected OW.

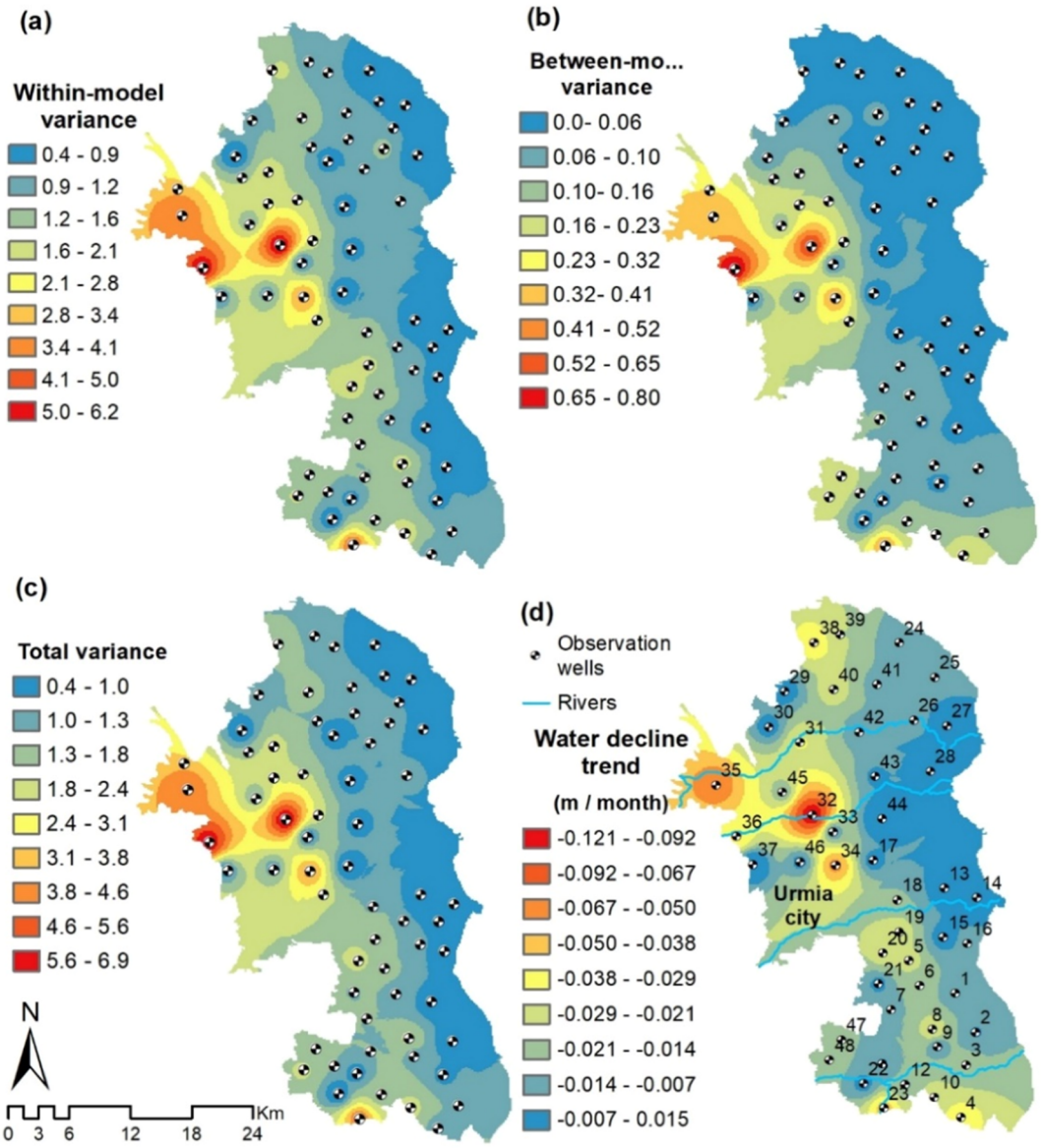

Fig. 9. Spatial distribution of variance of GWL by BMA: (a) within-model variance; (b) between-model variance; (c) total variance; (d) spatial distribution of the decline of water table.

are state variables at a set of particular OWs. Therefore, similarity between Fig. 9a and 9b stems from the variance of GWL time series, which are inevitably reflected in both between-model and within-model variances. Notably, the averaged value of between-model variance is affected by the GWL variances.

Scope of the Strategy based on BMA: The above results also show that the strategy is suitable as an aquifer management tool. Aquifer management in the Urmia aquifer up to 1990 was based on traditional common rules of the local communities, often with a high level of care for the wholesomeness of its water and equitable distributions among the users including tacitly meeting ecological and environmental needs. In less than one generation, the practice of pumpage undermined the long-established tradition. The model shows that it is quite feasible to develop a quantitative basis for management, in which the decline of water table can be arrested and a new regime of equitable distribution can be worked out by the full participation of the users and a representation of ecological and environmental needs.

The objectives of the study is not to develop a plan as a research work has no such mandates but to formulate desperately needed tools for planning practices in the future. The model developed here can now be applied in various ways. For instance, scenarios can be formulated to assess impacts of reduced recharge or increased abstraction. The immediate plan is to include the aquifer capacity as a variable in the model structure. This capacity, if feasible, would be a quick approach to assess the capacity of aquifers and set the alarms before depleting aquifers reach unrecoverable states.

Management of Urmia aquifer: Urmia aquifer supports the livelihood of a population of more than 1 million and is at the margin of the fertile crescent supporting the livelihood of its population for more than four millennia through the use of the traditional qanat system interwoven into the fabric of its cultural setting. Since 1990 there was a meteoric change in its irrigation culture as pumps became available. The aquifer was a common resource with little regulation to cope with new changes and therefore should sustainable practices not be developed for the aquifer, the emergence of “the tragedy of the commons” is inevitable and Fig. 9d shows that the process is already underway. This is supported by qualitative information obtained from a number of water users that since 2000, the experience of the decline of water table instigated the deepening of the wells. This impacted on water table with a further decline and created salinity problems. There is no study to

take a systemic view of the problems yet.

Whilst water resource planning with public participation is yet to be developed in the country, setting the equitable level of abstraction is haphazardly for not being evidence-based and thereby defining the level of unauthorised abstractions is a grey area. Nonetheless, a report by the West Azerbaijan Regional Water Authority (2016), attributes 80% of abstraction to unauthorised sourced. Based on the preliminary findings reflected by the paper, the emerging full picture is outlined as follows: (i) the decline in water table is not total yet as their sources of replenishment broadly remain unchanged and therefore the adverse situation can be arrested; (ii) any mitigation measure must be sustainable and for this a number of management plans are overdue including basin management plan for the Lake Urmia basin; aquifer management plans for each of the 12 aquifers surrounding Lake Urmia including Urmia aquifer and water cycle studies to ensure that water is equitably distributed among all sectors of water users; (iii) international experience shows that any plan without the participation of water users is likely to fail and therefore aquifer plans ought to define appropriate equitable levels; (iv) extension services are essential to train the water user communities with ways of efficient use of water.

6. Conclusion

A modelling strategy is presented to study Urmia aquifer, the West Azerbaijan province, northwest Iran. Due to the absence of any modern aquifer management practices, the aquifer is distressed as its water table is in decline based on 14 years of recorded water levels at 48 Observation Wells (OW). The modelling strategy for the sparse data incorporates three models, which comprise three AI models of Artificial Neural Network (ANN), Sugeno Fuzzy Logic (SFL) and Multiple-Neuro-Fuzzy (NF). A Bayesian Model Averaging (BMA) technique is employed to combine these modelling results, where BMA improves predictions and provides facilities for quantifying uncertainty.

The results of the three AI models show that there is no single model performing the best but they have convergences and divergences. BMA combines these modelling results into a single model, in which the combined model is a learning from the convergence and divergence of the AI models and as such it performs better than the individual models most of the time but overwhelmingly reduces the scatters in the error residuals. The paper refrains from ranking the models but recommends that the BMA medium is used for learning from models.

The analysis of the results shows that BMA is a suitable strategy for an operational aquifer management system and with a more monitoring of water table, it can respond to its inherent variations. Although Urmia aquifer is distressed and has come to this state in two decades or so, the paper argues that the aquifer is not dry yet and its rainfall patterns remain more or less the same and therefore its past healthy state can be reinstated. This will depend on developing appropriate defensible models, developing a series of management plans, giving high importance to inputs of opinions from the water users through their participation and never compromising the compensation flows for ecological and environmental functioning of the basin.

Acknowledgment

The authors would like to thank West Azerbaijan Regional Water Authority for their cooperation in data preparation. The project was supported financially by Iran National Science Foundation (96004114).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2019.02.011.

References

Alvisi, S., Mascellani, G., Franchini, M., Bardossy, A., 2006. Water level forecasting through fuzzy logic and artificial neural network approaches. Hydrol. Earth Syst. Sci. Discuss. 10 (1), 1–17.

Amirataee, B., Zeinalzadeh, K., 2016. Trends analysis of quantitative and qualitative changes in groundwater with considering the autocorrelation coefficients in west of Lake Urmia Iran. Environ. Earth Sci. 75 (5), 371.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. Artificial neural networks in hydrology. I: Preliminary concepts. Journal of Hydrologic Engineering, 5(2), pp. 115–123; 2000a.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. Artificial neural networks in hydrology. II: Hydrologic applications. Journal of Hydrologic Engineering, 5(2), pp. 124–137; 2000b.

Bazartseren, B., Hildebrandt, G., Holz, K.P., 2003. Short-term water level prediction using neural networks and neuro-fuzzy approach. Neurocomputing 55 (3–4), 439–450.

Berger, J.O., 1985. Statistical decision theory and Bayesian inference. Springer Verlag, (New York).

Bezdek, J.C., 1981. Objective function clustering. In *Pattern recognition with fuzzy objective function algorithms*. Springer, Boston, MA, pp. 43–93.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10 (2–3), 191–203.

Bisht, D.C.S., Raju, M., Joshi, M., 2009. Simulation of water table elevation fluctuation using fuzzy-logic and ANFIS. Comput. Model. New Technol. 13 (2), 16–23.

Chang, J., Wang, G., Mao, T., 2015. Simulation and prediction of suprapermafrost groundwater level variation in response to climate change using a neural network model. J. Hydrol. 529, 1211–1220.

Chen, C.H., Lin, Z.S., 2006. A committee machine with empirical formulas for permeability prediction. Comput. Geosci. 32 (4), 485–496.

Chen, M.S., Wang, S.W., 1999. Fuzzy clustering analysis for optimizing fuzzy membership functions. Fuzzy Sets Syst. 103 (2), 239–254.

Chiu, S.L., 1994. Fuzzy model identification based on cluster estimation. J. Intell. Fuzzy Syst. 2 (3), 267–278.

Chitsazan, N., Tsai, F.T.C., 2014. Uncertainty segregation and comparative evaluation in groundwater remediation designs: a chance-constrained hierarchical Bayesian model averaging approach. J. Water Resour. Plann. Manage. 141 (3), 04014061.

Chitsazan, N., Nadiri, A.A., Tsai, F.T.C., 2015. Prediction and structural uncertainty analyses of artificial neural networks using hierarchical Bayesian model averaging. J. Hydrol. 528, 52–62.

Chitsazan, N., Tsai, F.T.C., 2015. A hierarchical Bayesian model averaging framework for groundwater prediction under uncertainty. Groundwater 53 (2), 305–316.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. Int. J. Forecast. 5 (4), 559–583.

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: a review. J. Hydrol. 375 (3–4), 613–626.

Coppola Jr, E.A., Rana, A.J., Poulton, M.M., Szidarovszky, F., Uhl, V.W., 2005. A neural network model for predicting aquifer water level elevations. Groundwater 43 (2), 231–241.

Coulibaly, P., Anctil, F., Aravena, R., Bobée, B., 2001. Artificial neural network modeling of water table depth fluctuations. Water Resour. Res. 37 (4), 885–896.

Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., 2005. Groundwater level forecasting using artificial neural networks. J. Hydrol. 309 (1–4), 229–240.

Draper, D., 1995. Assessment and propagation of model uncertainty. J. Roy. Stat. Soc. Series B (Methodol.) 45–97.

Elshall, A.S., Tsai, F.T.C., 2014. Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm. J. Hydrol. 517, 105–119.

Emamgholizadeh, S., Moslemi, K., Karami, G., 2014. Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). Water Resour. Manage. 28 (15), 5433–5446.

Ghorbani, M.A., Khatibi, R., Karimi, V., Yaseen, Z.M., Zounemat-Kermani, M., 2018. "Learning from multiple models using artificial intelligence to improve model prediction accuracies: applications to river flows. J. Water Resour. Manage.

Gong, Y., Zhang, Y., Lan, S., Wang, H., 2016. A comparative study of artificial neural networks, support vector machines and adaptive neuro fuzzy inference system for forecasting groundwater levels near Lake Okeechobee Florida. Water Res. Manage. 30 (1), 375–391.

Hawkes, P., Khatibi, R., Sayers, P., 2005. Coastal flood forecasting: best practice in England and Wales. In: Coastal Engineering 2004, pp. 3036–3048.

Haykin, S., 1999. Simon. Neural networks: a comprehensive foundation, 2nd ed. Prentice Hall.

Heddam, S., Bermad, A., Dechemi, N., 2012. ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. Environ. Monit. Assess. 184 (4), 1953–1971.

Höting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statist. Sci. 382–401.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359–366.

Hsu, T.J., Sakakiyama, T., Liu, L.-F., 2002. A numerical model for wave motions and turbulence flows in front of a composite breakwater. Coastal Eng. 46 (1), 25–50. https://doi.org/10.1016/S0378-3839(02)00045-5.3.

Huang, C.N., Yu, C.C., 2016. Integration of Taguchi's method and multiple-input, multiple-output ANFIS inverse model for the optimal design of a water-cooled condenser. Appl. Therm. Eng. 98, 605–609.

Jalalkamali, A., Jalalkamali, N., 2011. Groundwater modeling using hybrid of artificial

neural network with genetic algorithm. Afr. J. Agric. Res. 6 (26), 5775–5784.
J.S.R. Jang, C.T. Sun, E. Mizutani. Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence; 1997.
Jeffreys, H., 1961. Theory of Probability. Clarendon Press, Oxford.
Kadkhodaie-Ilkhchi, A., Rezaee, M.R., Rahimpour-Bonab, H., Chehrazi, A., 2009. Petrophysical data prediction from seismic attributes using committee fuzzy inference system. Comput. Geosci. 35 (12), 2314–2330.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Statist. Assoc. 90 (430), 773–795.
Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. Environ. Modell. Software 23 (7), 835–845.
Khatibi, R., Gouldby, B., Sayers, P., McArthur, J., Roberts, I., Grime, A., Akhondi-asl, A., 2003. Improving coastal flood forecasting services of the Environment Agency. In: Proc. of the 1st International Conference on Coastal Management, Brighton, UK, pp. 70–82.
Khatibi, R., Ghorbani, M.A., Kashani, M.H., Kisi, O., 2011a. Comparison of three artificial intelligence techniques for discharge routing. J. Hydrol. 403 (3–4), 201–212.
Khatibi, R., Ghorbani, M.A., Aalami, M.T., Kocak, K., Makarynskyy, O., Makarynska, D., Aalinezhad, M., 2011b. Dynamics of hourly sea level at Hillarys Boat Harbour, Western Australia: a chaos theory perspective. Ocean Dyn. 61 (11), 1797–1807.
Khatibi, R., Naghipour, L., Ghorbani, M.A., Aalami, M.T., 2013. Predictability of relative humidity by two artificial intelligence techniques using noisy data from two Californian gauging stations. Neural Comput. Appl. 23 (7–8), 2241–2252.
Khatibi, R., Ghorbani, M.A., Naghipour, L., Jothiprakash, V., Fathima, T.A., Fazelifard, M.H., 2014. Inter-comparison of time series models of lake levels predicted by several modeling strategies. J. Hydrol. 511, 530–545.
Khatibi, R., Ghorbani, M.A., Pourhosseini, F.A., 2017. Stream flow predictions using nature-inspired firefly algorithms and a multiple model strategy–directions of innovation towards next generation practices. Adv. Eng. Inf. 34, 80–89.
Li, X., Tsai, F.T.C., 2009. Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod. Water Resour. Res. 45 (9).
Link, W.A., Barker, R.J., 2006. Model weights and the foundations of multimodel inference. Ecology 87 (10), 2626–2635.
Liu, P., Elshall, A.S., Ye, M., Beerli, P., Zeng, X., Lu, D., Tao, Y., 2016. Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. Water Resour. Res. 52 (2), 734–758.
Mohanty, S., Jha, M.K., Kumar, A., Sudheer, K.P., 2010. Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. Water Resour. Manage. 24 (9), 1845–1865.
Moosavi, V., Vafakhah, M., Shirmohammadi, B., Behnia, N., 2013. A wavelet-ANFIS hybrid model for groundwater level forecasting for different prediction periods. Water Resour. Manage. 27 (5), 1301–1321.
Nadiri, A.A., 2007. Water level evaluation in Tabriz underground area by artificial neural networks. University of Tabriz, Iran, MS Theses.
Nadiri, A.A., Fijani, E., Tsai, F.T.C., Moghaddam, A.A., 2013. Supervised committee machine with artificial intelligence for prediction of fluoride concentration. J. Hydroinform. 15 (4), 1474–1490.
Nadiri, A.A., Chitsazan, N., Tsai, F.T.C., Moghaddam, A.A., 2014. Bayesian artificial intelligence model averaging for hydraulic conductivity estimation. J. Hydrol. Eng. 19 (3), 520–532.
Nadiri, A.A., Gharekhani, M., Khatibi, R., Sadeghfam, S., Moghaddam, A.A., 2017. Groundwater vulnerability indices conditioned by supervised intelligence committee machine (SICM). Sci. Total Environ. 574, 691–706.
A.A. Nadiri, K. Naderi, R. Khatibi, M. Gharekhani. Modelling groundwater level variations by learning from multiple models using fuzzy logic, Hydrological Sciences, J. (just accepted for publication); 2018a, in press.
Nadiri, A.A., Sedghi, Z., Khatibi, R., Sadeghfam, S., 2018b. Mapping specific vulnerability of multiple confined and unconfined aquifers by using artificial intelligence to learn from multiple DRASTIC frameworks. J. Environ. Manage. 227, 415–428.
A.A. Nadiri, K. Naderi, R. Khatibi, M. Gharekhani. Modelling groundwater level variations by learning from multiple models using fuzzy logic; Hydrological sciences J. (https://doi.org/10.1080/02626667.2018.1554940); 2019, in press.
S.P. Newman, P.J. Wierenga. Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites; 2003.
Nie, S., Bian, J., Wan, H., Sun, X., Zhang, B., 2016. Simulation and uncertainty analysis for groundwater levels using radial basis function neural networks and support vector machines models. J. Water Supply: Res. Technol.-Aqua jws2016069.
Pulido-Calvo, I., Gutierrez-Estrada, J.C., 2009. Improved irrigation water demand forecasting using a soft-computing hybrid model. Biosyst. Eng. 102 (2), 202–218.
Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. 92 (437), 179–191.
Sadeghfam, S., Ehsanitabar, A., Khatibi, R., Daneshfaraz, R., 2018. Investigating 'risk'of groundwater drought occurrences by using reliability analysis. Ecol. Ind. 94, 170–184.
Saeed, R.A., Galybin, A.N., Popov, V., 2013. 3D fluid–structure modelling and vibration analysis for fault diagnosis of Francis turbine using multiple ANN and multiple ANFIS. Mech. Syst. Sig. Process. 34 (1–2), 259–276.
Singh, A., Mishra, S., Ruskauff, G., 2010. Model averaging techniques for quantifying conceptual model uncertainty. Groundwater 48 (5), 701–715.
Shiri, J., Kisi, O., Yoon, H., Lee, K.K., Nazemi, A.H., 2013. Predicting groundwater level fluctuations with meteorological effect implications—a comparative study among soft computing techniques. Comput. Geosci. 56, 32–44.
Sun, Y., Wendi, D., Kim, D.E., Liong, S.Y., 2016. Application of artificial neural networks in groundwater table forecasting–a case study in a Singapore swamp forest. Hydrol. Earth Syst. Sci. 20 (4), 1405–1412.
Szidarovszky, F., Coppola Jr, E.A., Long, J., Hall, A.D., Poulton, M.M., 2007. A hybrid artificial neural network-numerical model for ground water problems. Groundwater 45 (5), 590–600.
Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst., Man, Cyber. 1, 116–132.
Taormina, R., Chau, K.W., Sethi, R., 2012. Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. Eng. Appl. Artif. Intell. 25 (8), 1670–1676.
Tapoglou, E., Trichakis, I.C., Dokou, Z., Nikolos, I.K., Karatzas, G.P., 2014. Groundwater-level forecasting under climate change scenarios using an artificial neural network trained with particle swarm optimization. Hydrol. Sci. J. 59 (6), 1225–1239.
K.A. Tilford, K.J. Sene, R. Khatibi. Flood Forecasting Model Selection - A New Approach. In 'Flooding in Europe: Challenges and Developments in Flood Risk Management', Eds: S. Begum, J.W. Hall, M.J.F. Stive. Advances in Natural and Technological Hazards Research, Kluwer; 2005.
Trichakis, I.C., Nikolos, I.K., Karatzas, G.P., 2011. Artificial neural network (ANN) based modeling for karstic groundwater level simulation. Water Resour. Manage. 25 (4), 1143–1152.
Tsai, F.T.C., Li, X., 2008. Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. Water Resour. Res. 44 (9).
Tsai, F.T.C., 2010. Bayesian model averaging assessment on groundwater management under model structure uncertainty. Stoch. Env. Res. Risk Assess. 24 (6), 845–861.
Tsai, F.T.C., Elshall, A.S., 2013. Hierarchical Bayesian model averaging for hydrostratigraphic modeling: uncertainty segregation and comparative evaluation. Water Resour. Res. 49 (9), 5520–5536.
WAWA. West Azerbaijan Water Authority, provided the data through private communications; 2015.
Wöhling, T., Schöniger, A., Gayler, S., Nowak, W., 2015. Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. Water Resour. Res. 51 (4), 2825–2846.
Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour. Res. 40 (5).
Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8 (3), 338–353.