Science of the Total Environment, Volume 712, April 2020, Article number 135539
DOI:10.1016/j.scitotenv.2019.135539

Ensemble Modelling Framework for Groundwater Level Prediction in Urban Areas of India

[1]Basant Yadav, [2]Pankaj Kumar Gupta, [3]Nitesh Patidar and [4]Sushil Kumar Himanshu

[1]Postdoctoral Fellow in Rural Water Supply, Cranfield University
Cranfield Water Science Institute, Cranfield University, Vincent Building
Cranfield, Bedford, MK43 0AL
*Email ID: Basant.Yadav@cranfield.ac.uk

[2]Post-Doctoral Fellow, Faculty of Environment, University of Waterloo, Canada
Postal Address: 200 University Ave W, Waterloo, ON N2L 3G1
Email ID: pk3gupta@uwaterloo.ca

[3]Scientist 'B', Groundwater Hydrology Division
National Institute of Hydrology, Roorkee – 247667
Uttarakhand, India.
Email ID: niteshpatidar88@gmail.com

[4]Postdoctoral Fellow, Texas A&M Agrilife Research,
Texas A&M University System, Vernon, Texas, United States
Email ID: sushil.himanshu@ag.tamu.edu

Abstract: India is facing the worst water crisis in its history, and major Indian cities which accommodates about 50% of its population will be among highly groundwater stressed cities by 2020. In past few decades, the urban groundwater resources declined significantly due to over exploitation, urbanization, population growth and climate change. To understand the role of these variables on groundwater level fluctuation, we developed a machine learning based modelling approach considering singular spectrum analysis (SSA), mutual information (MI), genetic algorithm (GA), artificial neural network (ANN), and support vector machine (SVM). The developed approach was used to predict the groundwater levels in Bengaluru, a densely populated city with declining groundwater water resources. The input data consist of groundwater levels, rainfall, temperature, NOI, SOI, NIÑO3 and monthly population growth rate, and were pre-processed using mutual information, genetic algorithm and lag analysis. Later, the optimized input sets were used in ANN and SVM to predict monthly groundwater level fluctuations. The results suggest that the machine learning based approach with data pre-

processing predict groundwater levels accurately (R>85%). It is also evident from the results that the pre-processing techniques enhances the prediction accuracy and results were improved for 66% of the monitored wells. Analysis of various input parameters suggest inclusion of population growth rate is positively correlated with decrease in groundwater levels. The developed approach in this study for urban groundwater prediction can be useful particularly in cities where lack of pipeline/sewage/drainage lines leakage data hinders physical based modelling.

Keywords: Machine Learning, mutual information, genetic algorithm, artificial neural network, support vector machine, urbanization.

1. Introduction

Groundwater is an important fresh water resource for drinking, agricultural, and industrial purposes in many countries (Boulton and Hancock, 2006; Kulkarni et al., 2015; Mukherjee, 2018). Variation in groundwater levels are subjected differences between the supply and release of groundwater, gaining/loosing stream flow variations, tidal effects, urbanization, earthquake, land subsidence and meteorological phenomena as well as global climatic changes (Todd and Mays, 2005; Taylor et al., 2013; Fendorf and Benner, 2016; Levanon et al., 2017; Tang et al., 2017; Suryanarayana and Mahammood, 2019). A study conducted by Loáiciga (2003) concluded that the rise in groundwater use associated with predicted population growth would pose a higher threat to the aquifer than climate change. The groundwater level response to the hydro(geo)logical (such as groundwater recharge and discharge), meteorological (such as precipitation and temperature) and anthropogenic (such as urbanization and climate change) factors is highly nonlinear and complex (Khatri and Tyagi, 2015; Sapriza-Azuri et al., 2015; Zeng et al., 2017; Liu et al., 2018; Minnig et al., 2018). However, enhanced understanding of complex groundwater level response mechanism considering socio-hydrological heterogeneity is critical for sustainable planning and management of urban water supply (Barthel et al., 2016;

60 Rathnayaka 2016; Shekhar et al., 2013; Sekhar et al., 2018). In urban areas, water resources
61 management is of utmost importance as its dependency on groundwater for domestic and
62 industrial water supply is increasing due to rapid population growth, increasing per capita water
63 use and limited surface water from distant sources (Eckstein and Eckstein, 2003; Foster et al.,
64 2011).

65 The advanced numerical methods are capable in modelling complex groundwater flow
66 processes within a given domain. Groundwater level fluctuation has been modelled using
67 physical-based numerical models in several studies (Kim et al., 2008; Wang et al., 2008; Borsi
68 et al., 2013; Yousefi et al., 2019), however, large number of parameter to represent all the
69 physical processes makes groundwater simulations with physical based modelling complex.
70 Moreover, a reliable prediction of groundwater fluctuations requires physical properties of the
71 domain and model parameters to calibrate the model simulations.). Also, uncertainty
72 associated with hydrological, geological, topographical, meteorological and climatic data
73 makes numerical model calibration and validation challenging (Yoon et al., 2011; Barzegar et
74 al., 2017). Further, limitation of data availability, associated cost and time results in model
75 uncertainty and poor model performance (Kumar, 2015; Woodward et al., 2016; Valocchi et
76 al., 2017. Also, assumptions involved in solving the governing equations using the physical
77 based modelling makes the approach less competent in prediction as most of the variables (e.g.
78 groundwater level, evapotranspiration, rainfall) are less predictable. As a result, numerical
79 models tend to produce imperfect results in spite of the capturing the physical processes
80 successfully (Sun et al., 2016). Contrary to that, nonlinear based interdependencies feature of
81 machine learning based models overcomes the requirement of explicit characterization of the
82 physical properties, or accurate representation of physical parameters (Sahoo et al., 2017;
83 Yadav et. al., 2017) and need not to model the underlying physical processes. Over the last
84 decades, machine learning based models have been used in diverse research areas due to their

advantages over numerical models and have been proved efficient in capturing the complex physical processes especially in the data scarce regions (Yoon et al., 2011). However, both physical and data based modelling are based upon different philosophies complement each other with respect to their inherent strengths and limitations (Pandey et al., 2016).. While the important hydrological processes involved in a physically based model make up the black-box feature of a machine learning based model, the difficulty in accurate physical modelling can be alleviated by the powerful machine learning based models (Panda et al., 2010; Napolitano et al., 2010).

A review of literature reports a wide application of machine learning models in modelling nonlinear processes that are complex in nature (Chau et al. 2005; Sivapragasam et al. 2008). Artificial neural network (ANN) has been widely used in past decade for problems related to groundwater level predictions (Coppola et al., 2003; Coulibaly et al., 2001; Daliakopoulos et al., 2005; Nayak et al., 2006; Mohanty et al., 2010; Mohanty et al., 2015). Yoon et al., (2011) and Gong et al., (2016) used ANN in comparative studies while predicting the groundwater level fluctuations. ANN has been adopted by many researchers in the past to predict groundwater levels, however, its high sensitivity to the trained data, overfitting and dependency on hidden neurons are some major drawbacks (Hsu et al. 2002; Wu and Chau, 2011). Similarly, support vector machine (SVM) a relatively newer technique has also been used for the groundwater level prediction in various site conditions (Yoon et al., 2011; He et al., 2014; Gong et al., 2016; Zhou et al., 2017). More recently, fuzzy theory and genetic programming (GP) have also been used to study the groundwater levels (Kurtulus and Razack, 2010; Güler et al., 2012; Shiri and Kisi, 2011; Fallah-Mehdipour et al., 2013; Kasiviswanathan et al., 2016). Further, latest techniques like extreme learning machine (ELM) which are much simple in design and application then ANN or SVM have also been used in groundwater modelling studies (Yadav and Eliza, 2017; Alizamir et al., 2018). Further, Suryanarayana et

al., (2014) developed a wavelet (WA) based integrated WA-SVM model to predict monthly groundwater levels and their results suggest that WA-SVM performs better than auto regressive integrated moving average (ARIMA), ANN and SVM. Similarly, s study conducted by Sahoo et al., (2017) used a data pre-processing approach in developing a hybrid artificial neural network model (HANN) to predict seasonal groundwater level change in agricultural region of high plains and the Mississippi river valley alluvial aquifer.

Accuracy of machine leaning based groundwater level simulation or prediction, predominately depends on the type of input data used. It was pointed out in many studies that the generalization ability of machine learning based models are significantly influenced by selection of appropriate input variable (Maier and Dandy, 2000; Galelli et al., 2014; Quilty et al., 2016; Sahoo et al., 2017). The most obvious input variables in groundwater level predictions studies are rainfall, evaporation, temperature and pumping patterns (Yoon et al., 2011; Singh et al., 2014; Mohanty et al., 2015; Kasiviswanathan, 2016; Chang et al., 2016; Barzegar et al., 2017; Wunsch, 2018). Further, groundwater levels are also partially controlled by interannual to multidecadal climate variability (Kuss and Gurdak, 2014; Sahoo et al., 2017; Velasco et al., 2017). Therefore, appropriate input variable along with the application of pre-processing techniques has resulted in improved groundwater level prediction accuracy by capturing the seasonal variability and reducing the impact of noisy data (Wu et al., 2009; Wang et al., 2014; Sahoo et al., 2017). Overall, these studies have demonstrated the ability of data based modelling for developing a generalized non-linear relationship among hydro(geo)logical, meteorological and climatic input variables and groundwater.

In most studies, application of simple to complex groundwater models have been demonstrated in large agricultural catchments and there are very few studies predicting groundwater levels in urban areas (Coulibaly et al., 2001; Daliakopoulos et al., 2005; Wang et al., 2014; Shaoo et al., 2017; Wunsch et al., 2018; Yousefi et al., 2019). A study by Lerner (2002) describes how

135 urbanization affects groundwater cycle by way of changes to both the total water budget, and
136 to pathways recharge. Groundwater management in such circumstances is crucial to control the
137 decline groundwater levels and therefore, this requires a scientific understanding of urban
138 groundwater systems based on a hydrological, meteorological, climatological and
139 anthropogenic factors. In this study, we studied the impact of population growth, climatic
140 variability and hydro-meteorological variables on the groundwater level fluctuations by
141 combining the pre-processing techniques and advanced machine learning models. To study the
142 groundwater level fluctuation of a complex urban catchment, we have selected variables like
143 population growth rate (P), rainfall (R), temperature (T) and climatic variables. To the best
144 knowledge of the authors, this study is the first to couple SSA and SVM considering variables
145 like population growth rate to study an urban catchment. The novelty of the research work is
146 to study a highly urbanized catchment with limited groundwater resources using hydro-
147 meteorological, climatic and population data in hybrid SSA-MI-GA-ANN and SSA-MI-GA-
148 SVM models to predict the monthly groundwater level fluctuations. Henceforth in this article
149 'H' (SSA-MI-GA) will be used as a prefix in front of ANN and SVM to represent hybrid
150 models. The objectives of this research was to study the performance of original (ANN, SVM)
151 and hybrid (HANN, HSVM) models to predict groundwater levels of Bengaluru urban district
152 for one and two months ahead.

153 **2. Material and Methods**

154 **2.1 Study Area**

155 The study area (Fig. 1) is located in the south-eastern part of Karnataka and have geographical
156 extent of 2174 km^2. Total population of the area is 9.622 million (2011) with population density
157 of 4,378 people per km^2. The average annual rainfall of the area is 970 mm getting contribution
158 from the South-Western monsoon (54.18%), the North-Eastern monsoon (26.53%) the pre-
159 monsoon showers (18.53%) (Bengaluru water supply and sewerage board, Report, 2017). The
160 average maximum and minimum temperature is about 38˚C during summer and 15 ˚C during

winter, respectively. The relative humidity is about 86% during monsoon and 63% during dry months.

Physiography of the area comprises of rocky upland, plateau and flat topped hills (approx. 900m above mean sea level). Soils of the area could be categorised into red loamy soil and lateritic soil. This type of soil can be found in the eastern and southern part of study area which have hilly to undulating topography with granite and gneissic terrain. Further, lateritic soils are observed in western part of the study area where the terrain is undulating and gently sloping topography of peninsular gneissic region (Bengaluru water supply and sewerage board, Report, 2017). Groundwater occurs in phreatic conditions in weathered zones and under semi-confined to confined conditions in fractured and jointed rock formations.

Granites and Gneisses of peninsular gneissic group form the primary aquifers in the study area. Alluvium of thickness 20–25 m thick occur along the river courses possessing substantial groundwater potential. About 90% of groundwater structures tapping as shallow aquifers are yielding less than 1 litter per second. While, deep aquifers of yield ranged from 2 to 8 litter per second are located in parts of Bengaluru north and Anekal taluks. Transmissivity ranged from 10 to 280 m^2 /day (CGWB, 2008; Gulgundi and Shetty, 2018).

The study area contains Archean crystalline formation (Dharwar Province of the southern Indian peninsula) comprising peninsular gneiss's complex with small patch of hornblende schist in the northern part and intrusive closepet granites all along the western part (Chadwick et al. 1997; Mukherjee et al. 2018). The eastern edge of the Bengaluru city dominates to laterite of tertiary age, which occur as isolated patches capping crystalline rocks. Some small stretch of about 25km comprising unconsolidated sediments (Channapatna and Devanahalli) are found in study area.

Groundwater levels vary seasonally and found deepest during summer (April-May) and shallow during post monsoon (October-November). In general, decline in groundwater level

186 across the study area starts from late November and this falling trend in during post monsoon
187 can be attributed to erratic monsoon and rapid urbanization and land use change thus
188 minimizing groundwater recharge (Bengaluru water supply and sewerage board, Report,
189 2017).

190 The modelling approach using ANN, SVM, HANN and HSVM were assessed in
191 Bengaluru urban district to predict monthly groundwater levels changes. Monthly groundwater
192 level data of 2010 to 2017 (8 years) for 24 wells were acquired from the District Groundwater
193 Office, Groundwater Directorate Bengaluru, Karnataka. The selected wells are uniformly
194 distributed in the study area and represents the groundwater conditions across the district.
195 Monthly gridded rainfall data with a resolution of 0.25°×0.25° (Pai et al., 2014, 2015) and
196 average temperature time series with a resolution of 1°×1° (Srivastava et al., 2009) was also
197 collected for the period of 2010 to 2017. Further, the climatic parameters such as Southern
198 Oscillation Index (SOI), Northern Oscillation Index (NOI), and Niño3 were collected also
199 collected for each month for years 2010 to 2017 from National Oceanic and Atmospheric
200 Administration (NOAA, 2018a, 2018b and 2018c). SOI and NOI relate variability in the
201 atmospheric forcing of climate change in northern and southern mid-latitude hemisphere
202 regions and show interesting relationships in equatorial and extratropical teleconnections and
203 represent a wide range of local and remote climate signals (Schwing et al., 2002). Niño3 is an
204 index which is used to define El Niño and La Niña events covering large spatial area and with
205 different seasonal evolution. Apart from the climatic and hydro-meteorological data, we also
206 considered population growth rate to consider the impact of urbanization on groundwater
207 levels. The annual population growth rate data was obtained from World Population Review
208 (2018) and later was converted into monthly growth rate.
209

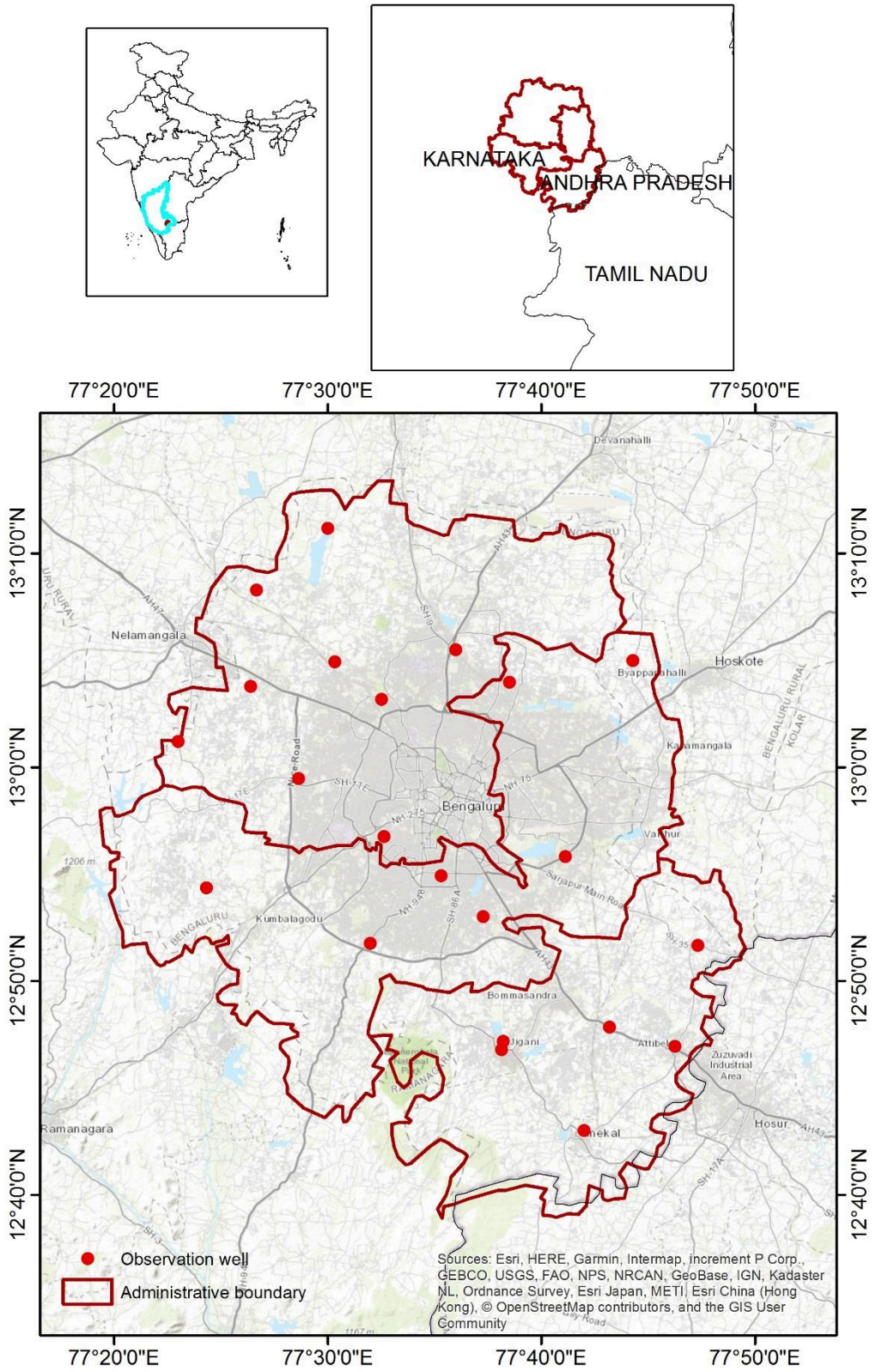


Fig. 1 Study area map of Bengaluru urban district with the monitoring well locations

2.2 Data processing and parametric optimization

The input data for the groundwater level prediction consist of groundwater levels, rainfall, temperature, NOI, SOI, NIÑO3 and monthly population growth rate. Data for the rainfall and temperature were processed and converted into time series. The input variables were pre-processed using singular spectrum analysis (SSA). Singular spectrum analysis decomposes the time series into a sum of a small number of interpretable components such as a slowly varying trend, oscillatory components and noise (Marques et al., 2006; Wang et al., 2015). SSA is a non-parametric technique of time series analysis based on principles of multivariate statistics (Vautard & Ghil, 1989; Dettinger et al., 1995). The given time series is decomposed into a set of independent time series to represent either a trend, periodic or quasi-periodic component or noise. As pointed out by Hanson et al., (2004) and applied by Sahoo et al. (2017) the variability of a hydrologic time series can be captured in first 10 reconstructed components (RCs) which are associated with the trend, oscillations or noise of the original time series. In this study, first 10 RCs of each input variable were extracted and later further processed using MI-GA approach to identify the most relevant RCs with respect to the groundwater levels.

The decomposed time series were further processed using mutual information and genetic algorithm to minimize the redundancy in the input data. Mutual information measure linear and non-linear dependencies between input and output variable. In this study, MI was used to establish nonlinear dependence between reconstructed components of the input variable and groundwater water levels. The interdependence between input and output variable using information theory obtained by measuring marginal entropy, conditional entropy and joint entropy. It takes a minimum value zero when there is no dependence between two variables, while a positive value suggest strong dependence among the considered input and output variables.

The entropy of a discrete random variable $x = (x_1, x_1 ... x_N)$ is denoted by $H(X)$. Where x_i refers to the possible values that X can take. $H(X)$ is defined as (Vergara and Estévez, 2014):

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 (p(x_i)) \quad (1)$$

where $p(x_i)$ is the probability mass function. For any two discrete random variable X and $Y = (y_1, y_1 ... y_M)$, the joint entropy is defined as (Vergara and Estévez, 2014):

$$H(X,Y) = -\sum_{j=1}^{M} \sum_{i=1}^{N} p(x_i, y_j) \log_2 (p(x_i, y_j)) \quad (2)$$

where $p(x_i, y_j)$ is the joint probability mass function of the variables X and Y. The conditional entropy of the variable X given Y is defined as (Vergara and Estévez, 2014):

$$H(Y \mid X) = -\sum_{j=1}^{M} \sum_{i=1}^{N} p(x_i, y_j) \log_2 (p(y_j \mid x_i)) \quad (3)$$

The joint entropy has values in the range,

$$\max(H(X), H(Y)) \le H(\{X,Y\}) \le H(X) + H(Y)$$

H (Y|X) is the amount of uncertainty left in Y when X is introduced, so it is less than or equal to the entropy of both variables, however it can be equal to the entropy if, the two variables have absolutely no dependencies.

Mutual information measures the level of interdependencies between two random variables. In case of feature selection, the approach is useful as it gives a quantifiable estimate of relevancy for a feature with respect to the output. Mutual information between two random variables is defined as (Vergara and Estévez, 2014):

$$I(X;Y) = \sum_{j=1}^{M} \sum_{i=1}^{N} p(x_i, y_j) . \log \left(\frac{p(x_i, y_j)}{p(x_i) . p(y_j)} \right) \quad (4)$$

where $I(X;Y)$ is mutual information between input X and output Y. For more information on information theory reader can refer to Vergara and Estévez (2014).

The relevancy of feature is counted if it provides information about output individually or together with other variables. However, the variable is counted redundant if it doesn't provide much information. Following the principle of maximum relevance (Eq. 5) and minimum redundancy (Eq. 6), most relevant RC of every input parameter was obtained using mutual information values and a genetic algorithm (Ludwig et al., 2009).

$$\mathrm{R}_{el} = \frac{1}{N}\sum_{i=1}^{N} I[X_i;Y], \tag{5}$$

$$R_{ed} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} I[X_i;X_j], \tag{6}$$

$$F_{\max} = R_{el} - R_{ed}, \tag{7}$$

where $\mathrm{R}_{el}\mathrm{V}$ represents average mutual information values showing the level of relevance between input (RCs) and output (groundwater levels) variables. R_{ed} represents average of all the mutual information values of the individual inputs. A fitness function $F_{\max}$ was solved using genetic algorithm to maximize the relevance and minimize the redundancy of the inputs.

In genetic algorithm variables are represented as chromosomes containing information about various decision variables that represent a decision or a solution. The fitness of randomly generated chromosomes is evaluated individually with respect to a target value. In this study, 40 randomly generated chromosomes were used for generation of 10 RCs. Objective function $F_{\max}$ was used to calculate the fitness of each chromosomes and later cross-over was performed, which introduces diversity in the population of the programs signifying the internal information exchange between the structures in the new and old population. The best model is identified by repeating the runs for a certain number of generations (80 in this study) or until a good

solution is obtained with fixed values of the controlling parameters. The optimization approach gave two best RCs that maximize the value of fitness function. Lastly, the pre-processed and optimized input variables were used in ANN and SVM to predict the monthly groundwater level.

3. Model Development

3.1 Artificial Neural Networks (ANN)

Artificial Neural Networks is a machine learning approach established as a robust tool for many hydrologic processes as simulation and prediction tool particularly when the underlying processes have complex nonlinear interrelationships (Govindaraju, 2000; Hsu et al., 2002). A common network of ANN comprises interconnected nodes called neurons arranged into input layer, hidden layer and output layer. The information is entered into data entry layer (input layer) which forwarded into hidden layer for the data processing and thereafter into output layer which generate the results for the given input (Dawson and Wilby, 1998). This type of network is called feed forward back propagation (FFBP) network in which the information passes only in forward direction from input layer, through hidden layer and finally to the output layer. The input vectors are $D \in R^n$ where $D = (X_1, X_2, ..., X_n)^T$, the outputs of N neurons in the hidden layer are $Z = (Z_1, Z_2, ..., Z_n)^T$ and the output from output layer are $Y \in R^m$ where $Y = (Y_1, Y_2, ..., Y_n)^T$. The weight and the threshold between the input layer and the hidden layers are w_{ij} and y_j, respectively. The following equations represent the neuron outputs in the hidden and output layer (Schalkoff, 1997):

$$Z_j = f\left(\sum_{i=1}^{n} w_{ij} X_i\right) \tag{8}$$

$$Y_k = f\left(\sum_{j=1}^{N} w_{kj} Z_j\right) \tag{9}$$

where a transfer function f is used to offer the rule for mapping the neuron's total input to its output.

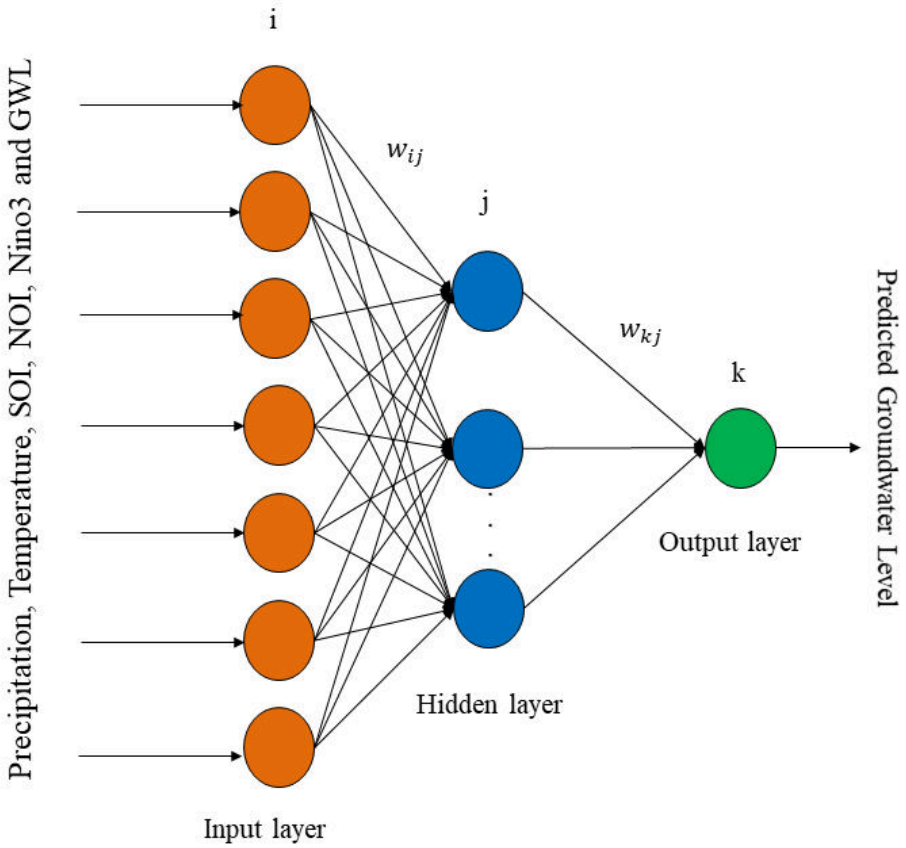


Fig. 2 A typical three-layer feed-forward ANN.

In this study, a three layer (input layer, hidden layer and output layer) ANN model (Fig 2) is developed to predic the groundwater level flectuations in Bengaluru urban district. The input and output data variables were divided into three groups training (70%), validation (15%) and testing(15%). The input layer consist of 7 neurons each for both models (ANN, HANN) and 1 output neuron (groundwater level). The developed model utilizes "newff" function to assign the initial weights randomly. An hyperbolic tangent sigmoid transfer function was used to process informtaion between input and hidden layer. Later, a linear transfer funtion "purelin" was used to transform the procssed infoirmation from hidden layer to the output layer (Schmid, 2009). The key parameters such as the learning rate and momentum in the network are obtained

by a trial and error procedure for each locations (wells). The model perfeormance was assessed using correlation coefficient (R), Root mean squared error (RMSE) and Normalised Mean Square Error (NMSE).

3.2 Support Vector Machine (SVM)

Support vector machine (SVM) is a machine learning approach proposed by Vapnik (1995) and as classification and regression procedure. SVM which based on the structural risk minimization principle has good generalization ability and is less prone to overfitting, (Vapnik and Vapnik, 1998; Vapnik, 2000; Yao et al., 2008). In regression problems, SVM uses kernel function to map the input vector in to a high dimensional feature space where the input vector is linearly separable (Wu et al., 2014). The developed SVM in this study was used to find a regression function that estimate the functional dependence between input and output variables $\{(x_1,t_1),...,(x_n,t_n)\}$. Where $x_i \in R^n$ in this study represents input variable (rainfall, temperature, past groundwater level, SOI, NOI and Niño3, population growth rate) while and $t_i \in R^n$ referred to as space of target output (groundwater level) value of n data lengths. SVM calculates the linear regression by solving (Vapnik, 1995) the following equations:

$$f(x) = w\varphi(x) + b \tag{10}$$

$$R_{SVM_s}(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n} L(x_i,t_i) \tag{11}$$

where $\varphi(x)$ is non-linear mapping function of x ; w is weight vector and b is a bias term; $C\frac{1}{n}\sum_{i=1}^{n} L(x_i,t_i)$ is the error component. To estimate the weight vector and bias, two positive slack variables ξ and ξ^* are added to limit the estimation error by the ε -insensitivity loss function (Vapnik and Vapnik, 1998) thus:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi + \xi^*\right) \tag{12}$$

$$\text{Subject to } \begin{cases} t_i - w\varphi(x_i) + b_i \le \varepsilon_i + \xi_i \\ w\varphi(x_i) + b_i - t_i \le \varepsilon_i + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1,\ldots, l \end{cases}$$

where C is the positive trade-off parameter for the degree of empirical error and l is the factor number in the training data.

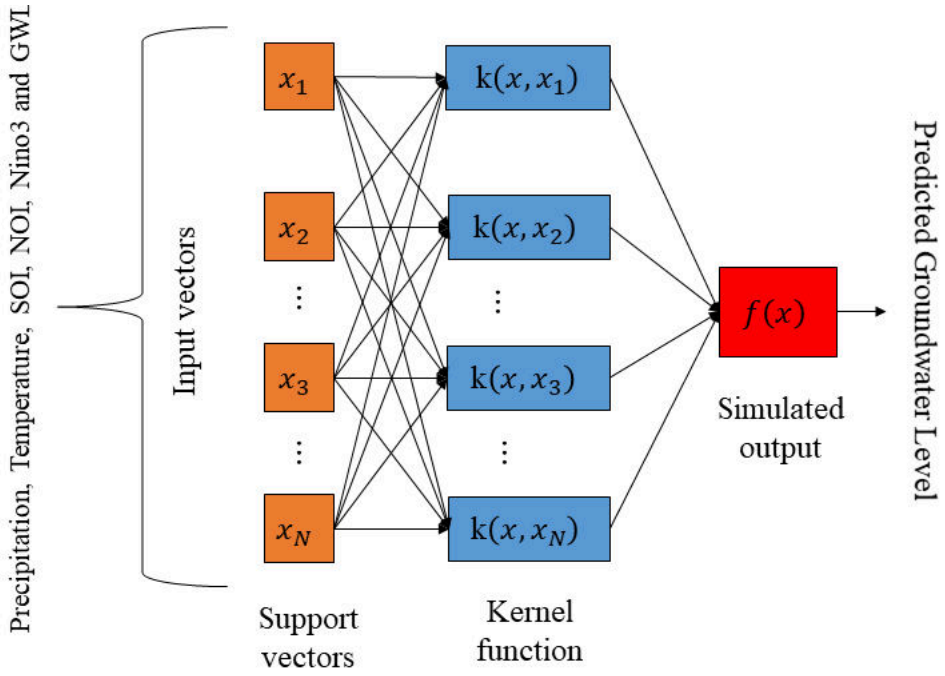


Fig. 3 The schematic representation of SVM

In previous hydrological prediction studies, radial basis functions (RBF) has been recommended as the most suited kernel function due to its capability to process highly complex parameter space (Rajasekaran et al., 2008; Yang et al., 2009; Wang et al., 2009; Yadav et al., 2016, Yadav and Eliza, 2017; Himanshu et al., 2017a; Yadav and Mathur, 2018). RBF is defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \quad (13)$$

where x_i and x_j are vectors in the input space, such as the vectors of features computed from training and testing. γ is defined by, $\gamma = -\frac{1}{2\sigma^2}$ for which σ is the Gaussian noise level of standard deviation. Figure 3 shows the schematic representation of SVM.

The SVM models for monthly groundwater level prediction using the pre-processed input data obtained from SSA and MI-GA approached was developed using LIBSVM toolbox. The model structure (Eq. 14 and 15) which were used in ANN, HANN model were kept same for SVM and HSVM as well. Prediction accuracy of SVM model depends on the suitable selection of kernels and parameters. It has been suggested in many studies that the radial basis function (RBF) performs well in hydrological forecasting problems and is hence it was considered in this study as well. The parameter of radial basis function was obtained using trial and error method for each well location. The developed models were used predict the monthly groundwater level using raw (SVM) pre-processed (HSVM) input variables for one and two months ahead.

3.3 The Ensemble Model

The ensemble prediction model using SSA, MI-GA, ANN and SVM was developed in MATLAB2014a. A separate code was written for SSA to decompose the input time series. MI-GA model was developed following the approach suggested by Ludwig et al., (2009). Further, ANN and SVM models were developed in MATLAB using 'newff' function and LIBSVM library, respectively. The input variable, monthly population growth rate, SOI, NOI, Niño3 were recorded as monthly time series. Temperature and rainfall data were extracted from the gridded (0.25°×0.25°) IMD data set and converted into monthly time series. The area covered under the gridded data for latitude 12.75° to 13.25° and Longitude 77.25° to 77.75° produced nine-time series for both temperature and rainfall, however, initial analysis revealed little

variability among them, therefore only one-time series for each temperature and rainfall was considered during groundwater level prediction. Lastly, the monthly groundwater level for the past two months was also added in the input data set. Each input variable was first decomposed using SSA and converted in 10 RCs. These RCs were supplied to an integrated MI-GA based model with the objective function to maximize the relevancy and minimize the redundancy resulted into two best RCs with respect to output variable (groundwater level). This procedure was repeated for each input variable and all the well locations. Once the optimized RCs with respect to each well water level was identified, the input-output combination was divided into training (70%), validation (15%) and testing (15%) data set. These combinations were used in both HANN and HSVM to predict the groundwater level at each location for one month and two months ahead. These hybrid models were later compared with the original model ANN and SVM in which raw input data was used for training, validation and testing. Figure. 4 depicts the procedure followed:

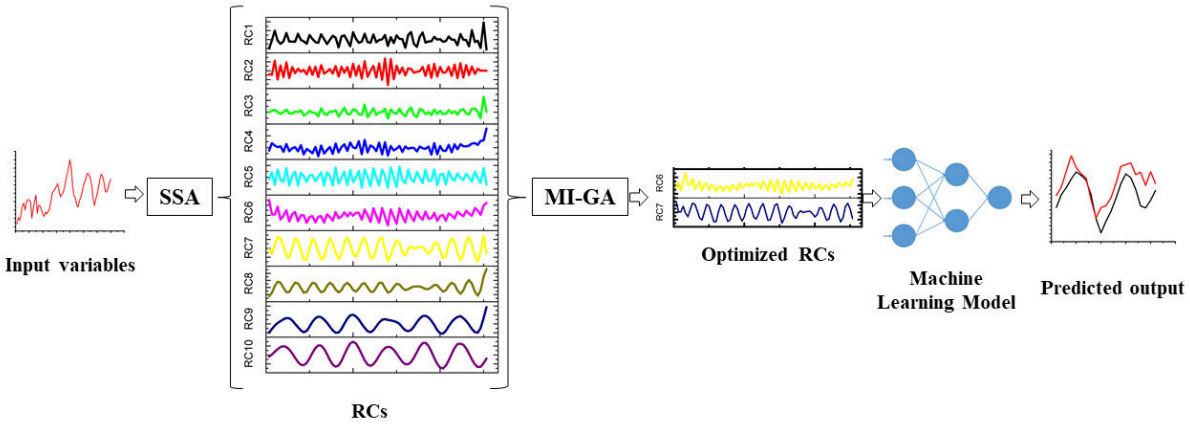


Fig. 4 Ensemble model using SSA, MI-GA and Machine learning models (ANN, SVM) to predict the groundwater levels.

4. Result and Discussion

4.1 Input selection and model development

In this study, a machine learning based approach considering data pre-processing techniques were developed to predict the groundwater level fluctuations in an urban area. The input data

395 consist of rainfall (R), temperature (T), SOI, NOI, Niño3, population growth rate (P) and
396 groundwater level (GWL). Temporal lag correlational analysis was preformed among the best
397 RC of groundwater and the best RC of population growth rate, southern oscillation index,
398 Northern oscillation index, Niño3, rainfall, temperature and past groundwater level,
399 respectively. Population growth rate was found to be strongly correlated with groundwater
400 levels at zero lag, which suggest that the increasing population and hence increased
401 groundwater abstraction had immediate impact on the groundwater level. Similarly,
402 temperature and Niño3 also showed strong correlation with groundwater level at zero-time lag,
403 however SOI time lag was found to be three months. Further, resulting time lags for NOI index
404 and R were two months. The obtained prediction model for one month (eq. 14) and two months
405 (eq 15) expressed as follows:
406
407 $$GWL_{t+1} = f(P_t, NOI_{t+2}, SOI_{t+3}, Nino3_t, T_t, R_{t+2}, GWL_t) \quad (14)$$
408 $$GWL_{t+2} = f(P_t, NOI_{t+2}, SOI_{t+3}, Nino3_t, T_t, R_{t+2}, GWL_{t+1}) \quad (15)$$
409 Later, all input variables were processed using SSA which generated 10 RCs of each input
410 variable. Subsequently, the obtained RCs were optimized using MI-GA approach
411 corresponding to each well locations. The optimized RCs of all the input variables were then
412 used in ANN and SVM model to predict the groundwater level for one and two months ahead.
413 ANN model was developed for each well location and hidden neurons were optimized using a
414 trial and error method. Hidden neurons play very significant role in the model prediction
415 accuracy and large number of input variable with high variability requires more hidden
416 neurons. The minimum hidden number 7 were obtained for Chikkabanavara well location,
417 however the highest number 21 was obtained at Thimmenahalli. Wells with the groundwater
418 level standard deviation more than 7 (Fig. 5) required 15 to 21 neurons. However, if the

groundwater level standard deviation varied between 1 to 5, resulting optimum neurons varied from 5 to 9.

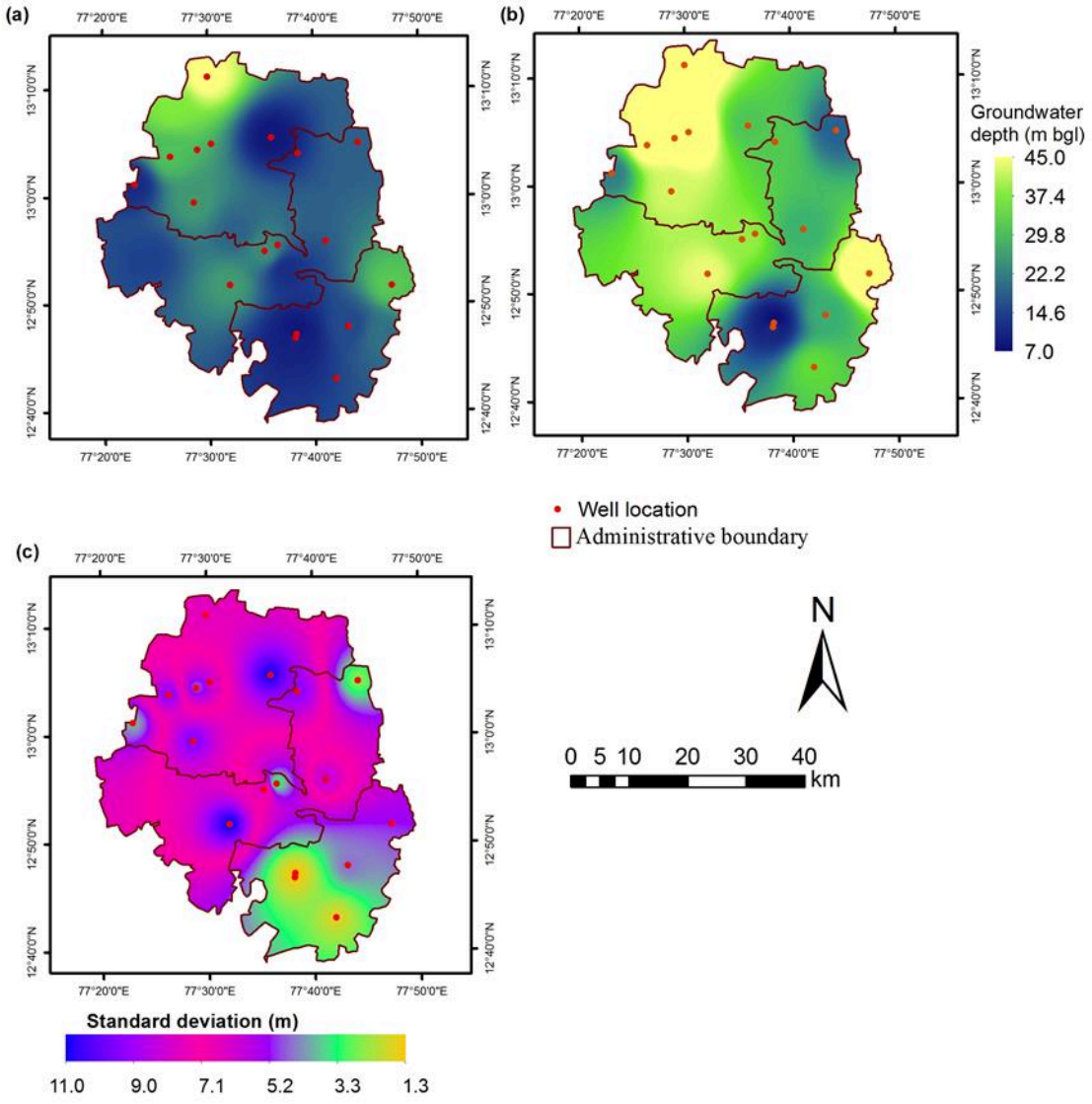


Fig. 5 Groundwater level during 2010 (a) and 2015 (b). Standard deviation in the groundwater level from 2010 to 2015 (c). A common colobar is shown for map (a) and (b) in the right.

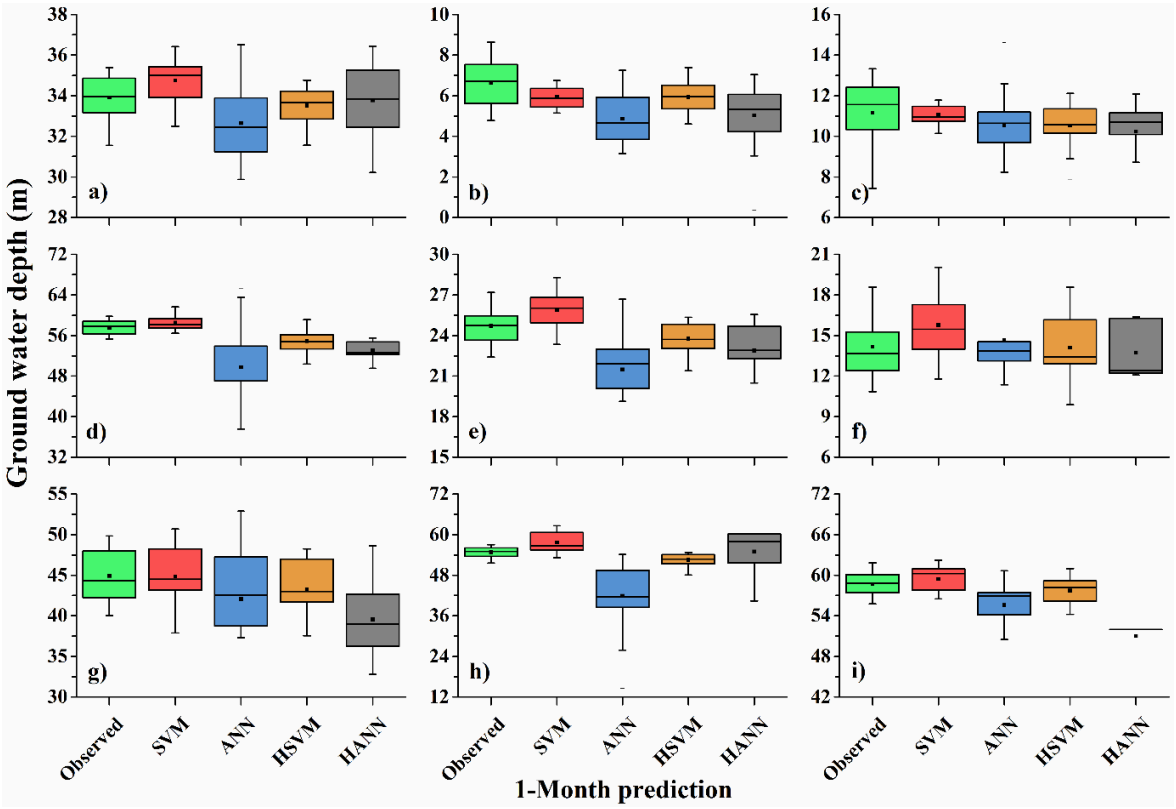
Similarly, SVM model parameters namely regularization constant, insensitive loss function and RBF parameter (γ) were also obtained corresponding to each well location. Regularization constant varied between 1.42 to 2.65 for all the location while insensitive loss function values between 0.032 to 0.098. Tuning of regularization constant is crucial for successful model development as it is a trade-off between the model complexity (flatness) and the degree to

which deviations larger than insensitive loss function are tolerated in optimization formulation. Value of C was kept low to achieve a balance output where the empirical error was minimized considering the model complexity. Insensitive loss function values were kept low for all the well locations as larger value can cause fewer selection of support vector which will result in less complex regression estimate. The parameter obtained during the model development (ANN and SVM) were kept similar for the prediction (1 month and 2 month) as well.

4.2 Groundwater level predicton

Long term simulations (2010-2015) were run to predict the 1-month and 2-months ahead groundwater depths for eighteen different locations across Bengaluru city in India. A substantial difference in predicted groundwater levels were observed under different machine learning approaches (Figures 6-9). Results presented here as box plots indicates the significant variability in simulated results across the simulation period under different machine learning approaches. The horizontal line and small solid square inside the box indicate the median and mean, respectively, and the ends of boxes indicate the 25^{th} and 75^{th} percentiles. Small line outside the boxes represent outliers or values greater than 1.5 interquartile ranges away from the 25th or 75th percentiles. In general, the simulated results were significantly improved under hybrid approaches (HANN and HSVM) as compared to conventional ANN and SVM approaches for both 1-month and 2-months ahead predictions. The results of the simulations are consistent with several other studies, which reported that the model predictability was further improved under hybrid approaches as compared to conventional machine learning approaches (Sahoo et al., 2017; Himanshu et al. 2017b; Yadav and Eliza, 2017). The results showed that 1-month ahead predictions were very precise and shows a better agreement with the observed groundwater level data. However, for 2-months ahead predictions, the predicted results were not very close to observed values, especially under conventional ANN and SVM approaches. Among HSVM and HANN, performance of HSVM was found better than HANN

455 at most of the locations for both 1-month and 2-months ahead predictions, except at Sarjapura,
456 Talaghattapura and Manduru locations for 2-months ahead predictions, where perforance of
457 HANN was found better (Figure 8d, 9d, 9g). In general, among all the four approaches,
458 performance of ANN was comparatively not found good for both 1-month and 2-months ahead
459 predictions at all the locations.



460
461 **Fig. 6:** 1-Month ahead predicted groundwater depths for locations a) Anekal, b) Jigani, c)
462 Bannerughatta, d) Sarjapura, e) Chandapura, f) Thimmenahalli, g) Byadarahalli, h)
463 Chikkabanavara, and i) Rajanukunte

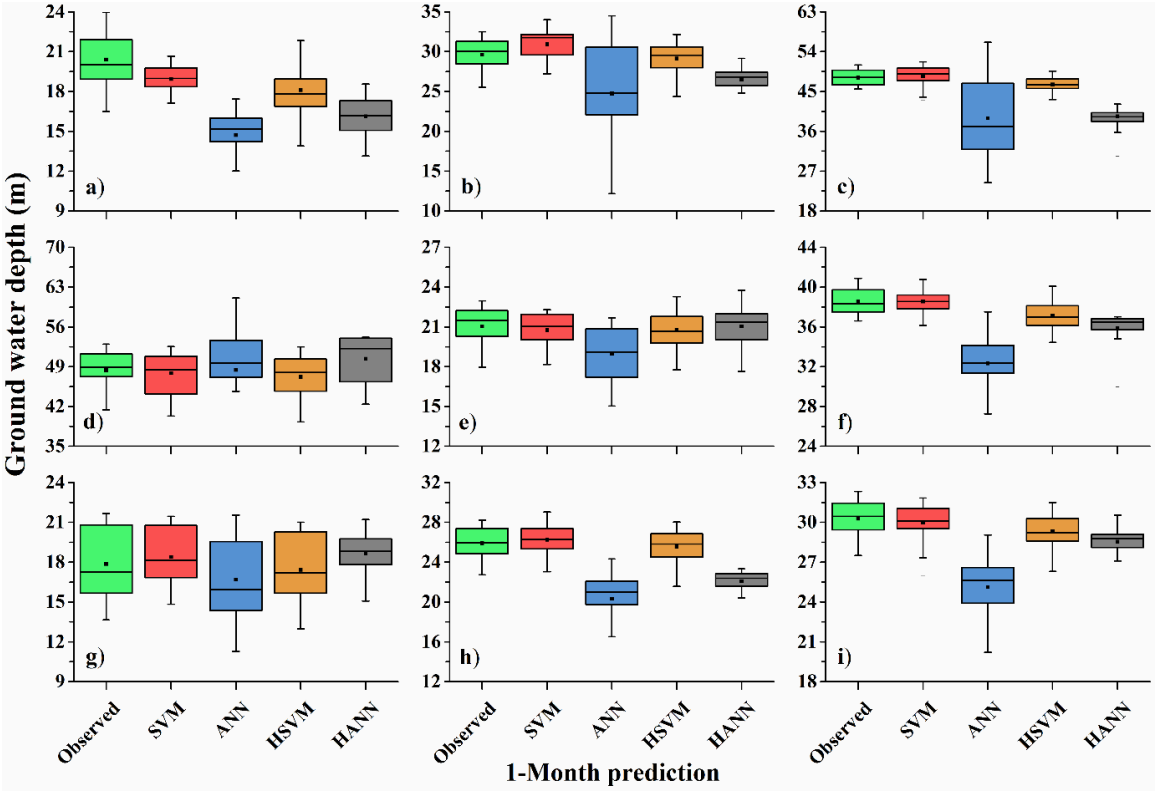


Fig. 7: 1-Month ahead predicted groundwater depths for locations a) Sondekoppa, b) Yelahanka, c) Adikemaranahalli, d) Talaghattapura, e) Tavarekere, f) Marenahalli, g) Manduru, h) Devarabeesanahalli, and i) K.Narayanapura

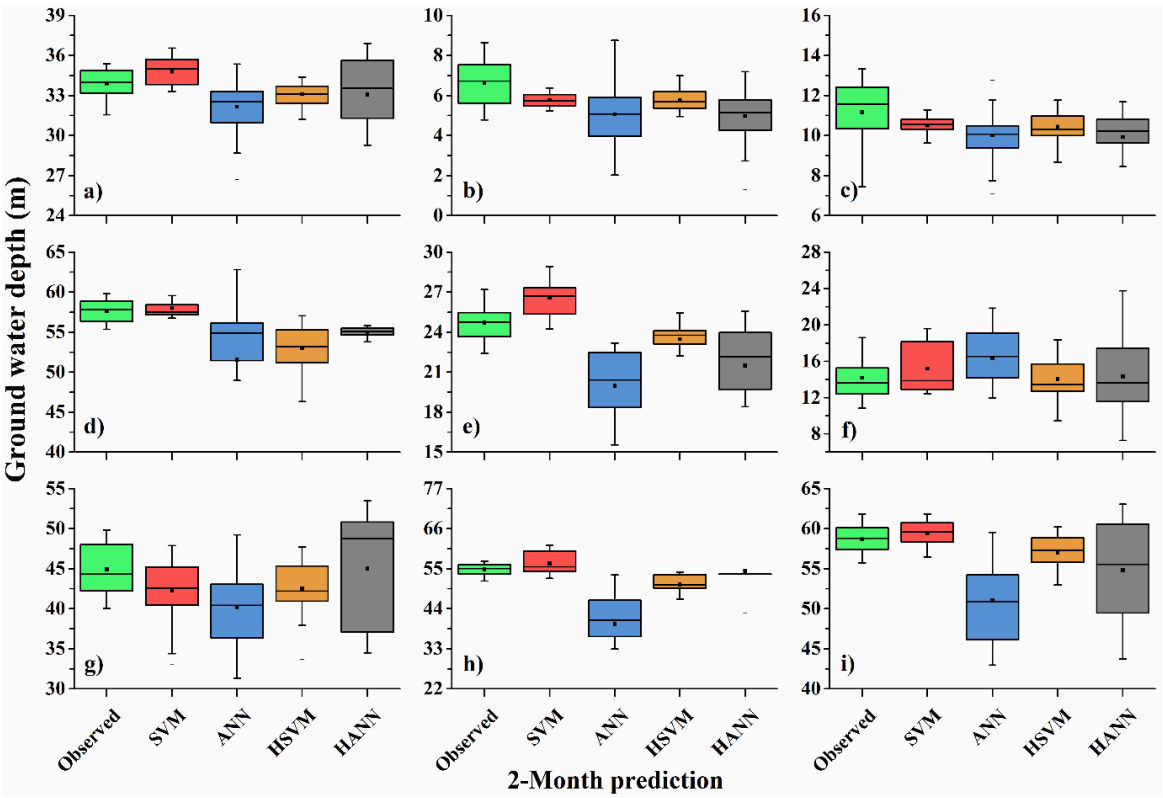


Fig. 8: 2-Month ahead predicted groundwater depths for locations a) Anekal, b) Jigani, c) Bannerughatta, d) Sarjapura, e) Chandapura, f) Thimmenahalli, g) Byadarahalli, h) Chikkabanavara, and i) Rajanukunte

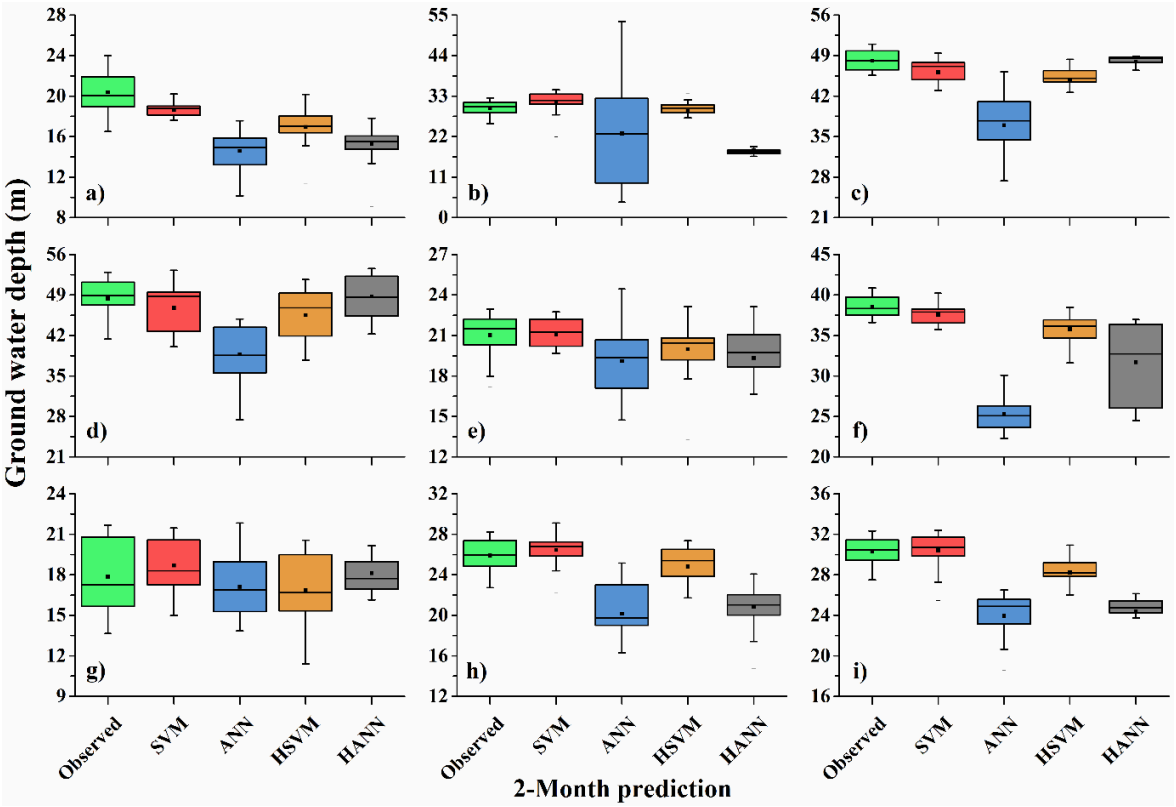


Fig. 9: 2-Month ahead predicted groundwater depths for locations a) Sondekoppa, b) Yelahanka, c) Adikemaranahalli, d) Talaghattapura, e) Tavarekere, f) Marenahalli, g) Manduru, h) Devarabeesanahalli, and i) K.Narayanapura

Table 1 show the average statistics of model performance during 1 and 2 months ahead groundwater level prediction. The results show that the performance of ANN model was improved significantly when used with pre-processed data (HANN). Performance statistics of ANN during 1 month ahead prediction improved approximately by 55%, 35% and 64% for R, RMSE and NMSE, respectively. In case of SVM, the improvement was 1.85%, 0.36% and 17% for same statistical indicators. Similar improvements were observed when the models were used for 2 months ahead prediction, the performance of ANN improved upto 50%, 31% and 56 % for R, RMSE and NMSE, respectively. However, the improvement in case of SVM was 0.97%, 3.77% and 11% for R, RMSE and NMSE, respectively. HSVM is the superior method to predict groundwater fluctuations throughout the study area and fortifies that model could map these complex interdependences of climatic variations, population growths on groundwater level fluctuations. These results indicate that the HSVM and SVM model of this

488 study is more likely to learn the complex relationship of groundwater fluctuation with urban
489 environment for the given data than the ANN. The evaluation of the proposed technique in an
490 urban area having a complex hydrological regime shows that the technique provides improved
491 results for groundwater level prediction when compared to traditional techniques.

492 Table 1 Original and hybrid model comparison for one and two month prediction using
493 average statistics for 19 well locations

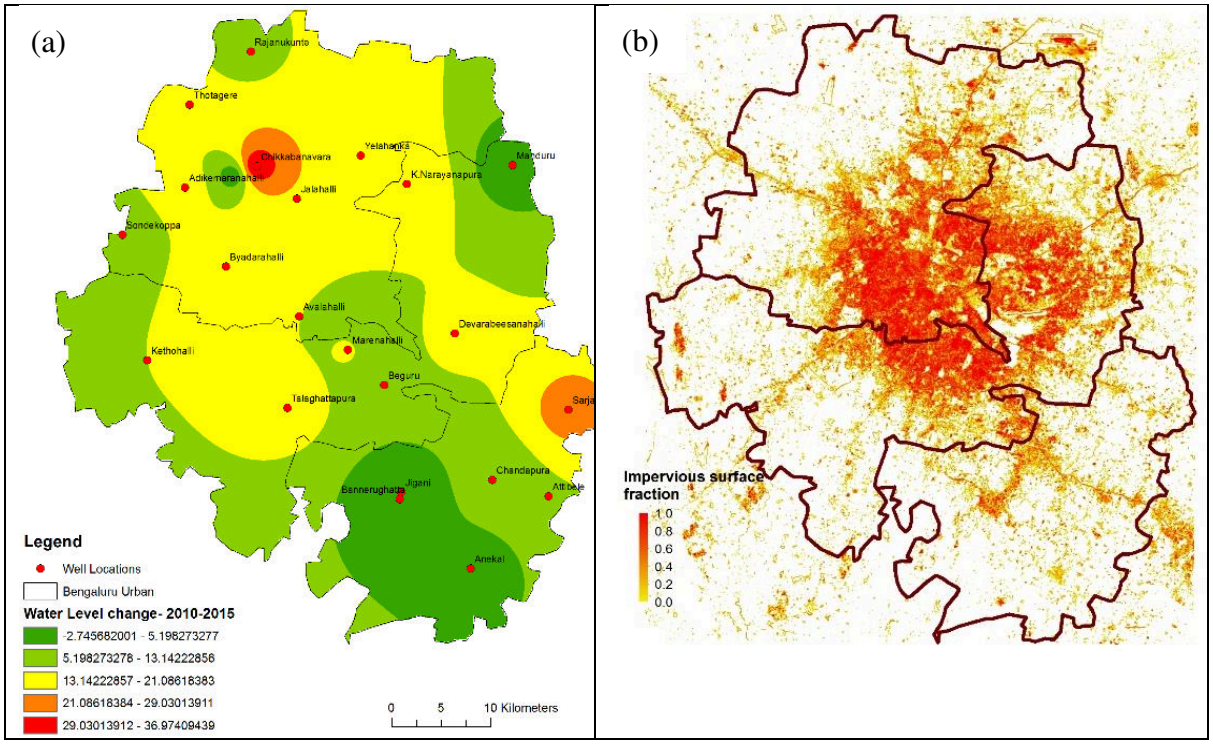
1 Month ahead prediction				
	ANN	SVM	HANN	HSVM
R	0.22	0.862	0.492	0.88
RMSE	6.01	1.361	3.861	1.36
NMSE	16.97	0.83	5.96	0.68
2 months ahead prediction				
R	0.14	0.71	0.29	0.72
RMSE	7.49	2.01	5.16	1.94
NMSE	27.00	1.32	11.63	1.17

494

495 **4.3 Groundwater Condition with Urbanization**

496 Change in groundwater level was investigated utilizing water level record of 22 wells during
497 2010 and 2015 which shows considerable decline in the levels (Fig. 10 (a)). Groundwater
498 decline in the study area ranges from 5 to 37 m bgl (below ground level) where the decline is
499 highest at Chikkabanavar and Sarjapura locations, as highlighted in red color on the map.
500 Whereas, increase in water levels were also observed at some locations in the Southern part of
501 the study area. The urbanization could be a major cause behind the depleting groundwater
502 mainly because urbanization has led to incresed groundwater pumping due to increased water
503 demand. Although the water needs of the city is mainly met by surface water imported from
504 the Cauvery River, it has been unable to catch up with the increasing water demand due to
505 rapid population growth and expansion of the city. As a result, groundwater satisfies a large

506 proportion of current water need. Further, increased impervious surface, which is associated
507 with urbanization, has led to decreased groundwater recharge in the area. Figure 10 (b) shows
508 the impervious surface fractions within 30 m × 30 m grids derived using Landsat image (Fig.
509 10 (b)). The comparison of groundwater change map and the impervious surface map shows
510 that the higher declines are not at the urban centre, but towards the urban periphery. This is
511 mainly due to the fact the groundwater pumping is happening mainly at peri-urban areas
512 (Sekhar et al., 2018). This also indicates that the dominating factor behind the decline in the
513 study area is pumping as compared to decreased groundwater recharge due to urbanization.
514 Furthermore, the decreased recharge could have been componsated by the leaking pipes and
515 wastewater to some extent in the study area. The competitive demands of water from various
516 sectors put additional pressure on groundwater resources. Kapetas et al. (2019) found that the
517 increasing competition for water has led to a water deficit in the agricultural sector, an unmet
518 environmental flow and a reduced capacity for urban supply during drought conditions.
519 Therefore, to manage water resources where competitive demands co-exist, coordinated multi-
520 level institutional relationships are important to improve water management practices and
521 water allocations (Kapetas et al., 2019).



522

523 **Fig. 10** Groundwater change map overlaid with observation wells (*a*) and impervious surface
524 map (*b*). The values in map (*b*) shows the fraction of impervious surface within the 30 m × 30
525 m grids for the year 2010.

526 Though the hybrid HANN and HSVM models predicts the groundwater level fluctuations
527 accurately, it does not give information about the corresponding physical processes in the
528 aquifer. Further, the impact of additional input variables such as groundwater recharge from
529 pipes leakage/sewers/drains has not been included in this study which could help in reducing
530 the prediction error. Therefore, there are environments and applications for which each model
531 type excels. In future studies, the methodology can be improved further by considering more
532 advanced and efficient machine learning techniques, like deep algorithms, random forest, etc.
533 Further, there is a scope in optimizing the number of parameters using other advanced
534 optimization approaches.

535

536

5. Conclusion

In this study, we proposed an ensemble machine learning based modelling approach using input data pre-proceesing for prediction of grounwater level fluctuations. The developed approach was applied and assesed for monthly groundwater level prediction at an densly populated urban city (Bengaluru) in India. The selected area has been under severe water stress and groundwater table has declined significantly in last 10 years due to urbanization and heavy pumping. The hybird models (HANN and HSVM) perform better than the orginal models (ANN and SVM) while predicting grounwater level fluctuations. It was also found that prediction accuracy decreases as we increase the time lead for both orginal and hybird models.

It is intresting to observe that the population growth rate provided positive information and captured the impact of urbanization on the groundwater level flectuations. Further, analysis of groundwater level decline (2010-2015) along with impervious surface suggest that the reason for declining trend in groundwater of urban centers is increased groundwater pumping rather than the decreased groundwater recharge due to urbanization. The results obtained from this study would be useful in indetifying the causes for groundwater decline in urban centers under various climatic conditions. The developed appraoch would be useful perticularly in the urban areas where physical based modelling is challenging due to scarcity of pipeline leakage or sweage/drainage line leakage data.

Acknowledgement

This study was supported by National Postdoctoral Fellowship (NPDF) grant (PDF/2017/000415) funded by Science and Engineering Research Board (SERB), India. The authors would like to acknowledge the District Groundwater Office, Groundwater Directorate Bengaluru, Karnataka for supplying the data.

Conflict of Interest

None

Data avialbility statement

The data used in the study can be provided on request by first author.

References

Alizamir, M., Kisi, O., and Zounemat-Kermani, M. (2018). Modelling long-term groundwater fluctuations by extreme learning machine using hydro-climatic data. *Hydrological Sciences Journal*, *63*(1), 63-73.

Bengaluru water supply and sewerage board (2017). Bengaluru water supply and Sewerage project (phase 3) [Available at http://open_jicareport.jica.go.jp/pdf/12300356_01.pdf]

Barthel, R., Ziller, R., Leinberger, A., and Hörhan, T. (2016). Changes to the quantitative status of groundwater and the water supply. In Regional Assessment of Global Change Impacts (pp. 561-567). Springer, Cham.

Barzegar, R., Fijani, E., Moghaddam, A. A., and Tziritis, E. (2017). Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models. *Science of the Total Environment*, *599*, 20-31.

Borsi, I., Rossetto, R., Schifani, C., and Hill, M. C. (2013). Modeling unsaturated zone flow and runoff processes by integrating MODFLOW-LGR and VSF, and creating the new CFL package. *Journal of Hydrology*, *488*, 33-47.

Boulton, A.J., and Hancock, P.J. (2006). Rivers as groundwater-dependent ecosystems: a review of degrees of dependency, riverine processes and management implications. *Australian Journal of Botany*, *54*(2), pp.133-144.

Chadwick, B., Vasudev, V. N., and Hegde, G. V. (1997). The Dharwar craton, southern India, and its Late Archaean plate tectonic setting: current interpretations and controversies. *Proceedings of the Indian Academy of Sciences-Earth and Planetary Sciences*, *106*(4), 249-258.

Chang, F. J., Chang, L. C., Huang, C. W., and Kao, I. F. (2016). Prediction of monthly regional groundwater levels through hybrid soft-computing techniques. *Journal of Hydrology*, *541*, 965-976.

Chang, J., Wang, G. and Mao, T., (2015). Simulation and prediction of suprapermafrost groundwater level variation in response to climate change using a neural network model. *Journal of Hydrology*, *529*, pp.1211-1220.

Chau, K. W., Wu, C. L., and Li, Y. S. (2005). Comparison of several flood forecasting models in Yangtze River. *Journal of Hydrologic Engineering*, *10*(6), 485-491.

Coppola Jr, E., Szidarovszky, F., Poulton, M., and Charles, E. (2003). Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions. *Journal of Hydrologic Engineering*, *8*(6), 348-360.

Coulibaly, P., Anctil, F., Aravena, R., and Bobée, B. (2001). Artificial neural network modeling of water table depth fluctuations. *Water Resources Research*, *37*(4), 885-896.

Daliakopoulos, I. N., Coulibaly, P., and Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*, *309*(1-4), 229-240.

Dawson, C. W., and Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, *43*(1), 47-66.

Dettinger, M. D., Ghil, M., Strong, C. M., Weibel, W., and Yiou, P. (1995). Software expedites singular-spectrum analysis of noisy time series. *EOS, Transactions American Geophysical Union*, *76*(2), 12-21.

Eckstein, G. E., and Eckstein, Y. (2003). A hydrogeological approach to transboundary ground water resources and international law. *Am. U. Int'l L. Rev.*, *19*, 201.

Fallah-Mehdipour, E., Haddad, O. B., and Mariño, M. A. (2013). Prediction and simulation of monthly groundwater levels by genetic programming. *Journal of Hydro-Environment Research*, *7*(4), 253-260.

Fendorf, S., and Benner, S. G. (2016). Hydrology: Indo-Gangetic groundwater threat. *Nature Geoscience*, *9*(10), 732.

Foster, S., Hirata, R., Misra, S., and Garduno, H. (2011). Urban groundwater use policy: balancing the benefits and risks in developing nations. GW-MATE Strategic Overview Series 3. World Bank, Washington, DC.

Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., and Gibbs, M. S. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models. *Environmental Modelling & Software*, *62*, 33-51.

Gong, Y., Zhang, Y., Lan, S., and 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 Resources Management*, *30*(1), 375-391.

Govindaraju, R. S. (2000). Artificial neural networks in hydrology. II: hydrologic applications. Journal of Hydrologic Engineering, 5(2), 124-137.

Güler, C., Kurt, M. A., Alpaslan, M., and Akbulut, C. (2012). Assessment of the impact of anthropogenic activities on the groundwater hydrology and chemistry in Tarsus coastal plain (Mersin, SE Turkey) using fuzzy clustering, multivariate statistics and GIS techniques. *Journal of Hydrology*, *414*, 435-451.

Gulgundi, M. S., and Shetty, A. (2018). Groundwater quality assessment of urban Bengaluru using multivariate statistical techniques. *Applied water science*, *8*(1), 43.

Hanson, R. T., Newhouse, M. W., and Dettinger, M. D. (2004). A methodology to asess relations between climatic variability and variations in hydrologic time series in the southwestern United States. *Journal of Hydrology*, *287*(1-4), 252-269.

He, Z., Wen, X., Liu, H., and Du, J. (2014). A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology*, *509*, 379-386.

Himanshu, S. K., Pandey, A., and Yadav, B. (2017). Assessing the applicability of TMPA-3B42V7 precipitation dataset in wavelet-support vector machine approach for suspended sediment load prediction. *Journal of Hydrology*, *550*, 103-117.

Himanshu, S. K., Pandey, A., & Yadav, B. (2017b). Ensemble wavelet-support vector machine approach for prediction of suspended sediment load using hydrometeorological data. Journal of Hydrologic Engineering, 22(7), 05017006.

Hsu, K. L., Gupta, H. V., Gao, X., Sorooshian, S., and Imam, B. (2002). Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis. *Water Resources Research*, *38*(12), 38-1.

Kapetas, L., Kazakis, N., Voudouris, K., and McNicholl, D. (2019). Water allocation and governance in multi-stakeholder environments: Insight from Axios Delta, Greece. *Science of the Total Environment*, 695, 133831.

Kasiviswanathan, K. S., Saravanan, S., Balamurugan, M., and Saravanan, K. (2016). Genetic programming based monthly groundwater level forecast models with uncertainty quantification. *Modeling Earth Systems and Environment*, *2*(1), 27.

Khatri, N., and Tyagi, S. (2015). Influences of natural and anthropogenic factors on surface and groundwater quality in rural and urban areas. *Frontiers in Life Science*, *8*(1), pp.23-39.

Kim, N. W., Chung, I. M., Won, Y. S., and Arnold, J. G. (2008). Development and application of the integrated SWAT–MODFLOW model. *Journal of Hydrology*, *356*(1-2), 1-16.

Kulkarni, H., Shah, M., and Shankar, P.V. (2015). Shaping the contours of groundwater governance in India. *Journal of Hydrology: Regional Studies*, *4*, pp.172-192.

Kumar, C.P., (2015). Modelling of groundwater flow and data requirements. International Journal of Modern Sciences and Engineering Technology, 2(2), 18-27.

Kurtulus, B., and Razack, M. (2010). Modeling daily discharge responses of a large karstic aquifer using soft computing methods: artificial neural network and neuro-fuzzy. *Journal of Hydrology*, *381*(1-2), 101-111.

Kuss, A. J. M., and Gurdak, J. J. (2014). Groundwater level response in US principal aquifers to ENSO, NAO, PDO, and AMO. *Journal of Hydrology*, *519*, 1939-1952.

Lerner, D. N. (2002). Identifying and quantifying urban recharge: a review. *Hydrogeology journal*, *10*(1), 143-152.

Levanon, E., Yechieli, Y., Gvirtzman, H., and Shalev, E. (2017). Tide-induced fluctuations of salinity and groundwater level in unconfined aquifers–Field measurements and numerical model. *Journal of hydrology*, *551*, pp.665-675.

674 Liu, C.Y., Chia, Y., Chuang, P.Y., Chiu, Y.C. and Tseng, T.L. (2018). Impacts of
675 hydrogeological characteristics on groundwater-level changes induced by
676 earthquakes. *Hydrogeology journal*, *26*(2), 451-465

677 Loáiciga, H. A. (2003). Climate change and ground water. *Annals of the Association of*
678 *American Geographers*, *93*(1), 30-41.

679 Ludwig Jr, O., Nunes, U., Araújo, R., Schnitman, L., and Lepikson, H. A. (2009). Applications
680 of information theory, genetic algorithms, and neural models to predict oil
681 flow. *Communications in Nonlinear Science and Numerical Simulation*, *14*(7), 2870-2885.

682 Maier, H. R., and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of
683 water resources variables: a review of modelling issues and applications. *Environmental*
684 *Modelling & Software*, *15*(1), 101-124.

685 Marques, C. A. F., Ferreira, J. A., Rocha, A., Castanheira, J. M., Melo-Gonçalves, P., Vaz, N.,
686 and Dias, J. M. (2006). Singular spectrum analysis and forecasting of hydrological time
687 series. *Physics and Chemistry of the Earth, Parts A/B/C*, *31*(18), 1172-1179.

688 Minnig, M., Moeck, C., Radny, D. and Schirmer, M., (2018). Impact of urbanization on
689 groundwater recharge rates in Dübendorf, Switzerland. *Journal of hydrology*, *563*, 1135-1146.

690 Mohanty, S., Jha, M. K., Kumar, A., and Sudheer, K. P. (2010). Artificial neural network
691 modeling for groundwater level forecasting in a river island of eastern India. *Water Resources*
692 *Management*, *24*(9), 1845-1865.

693 Mohanty, S., Jha, M. K., Raul, S. K., Panda, R. K., and Sudheer, K. P. (2015). Using artificial
694 neural network approach for simultaneous forecasting of weekly groundwater levels at multiple
695 sites. *Water Resources Management*, *29*(15), 5521-5532.

696 Mukherjee, A. (2018). *Groundwater of South Asia*. Springer.

697 Mukherjee, S., Ghosh, G., Das, K., Bose, S., and Hayasaka, Y. (2018). Geochronological and
698 geochemical signatures of the granitic rocks emplaced at the north-eastern fringe of the Eastern
699 Dharwar Craton, South India: Implications for late Archean crustal growth. *Geological*
700 *Journal*, *53*(5), 1781-1801.

701 Napolitano, G., See, L., Calvo, B., Savi, F., and Heppenstall, A. (2010). A conceptual and
702 neural network model for real-time flood forecasting of the Tiber River in Rome. Physics and
703 Chemistry of the Earth, Parts A/B/C, 35(3-5), 187-194.

704 Nayak, P. C., Rao, Y. S., and Sudheer, K. P. (2006). Groundwater level forecasting in a shallow
705 aquifer using artificial neural network approach. *Water Resources Management*, *20*(1), 77-90.

706 NOAA (2018b), Northern Oscillation Index (NOI) Data, Earth Syst. Res. Lab., Phys. Sci. Div.,
707 Boulder, Colo. [Available at: https://www.esrl.noaa.gov/psd/data/correlation/noi.data]

708 NOAA (2018c), Eastern Tropical Pacific SST (NIÑO3) Data, Earth Syst. Res. Lab., Phys. Sci.
709 Div., Boulder, Colo. [Available at: https://www.esrl.noaa.gov/psd/data/correlation/nina3.data]

NOAA(2018a), Southern Oscillation Index (SOI) Data, Earth Syst. Res. Lab., Phys. Sci. Div., Boulder, Colo. [Available at: https://www.esrl.noaa.gov/psd/data/correlation/soi.data]

Pai, D.S., Sridhar, L., Badwaik, M.R. and Rajeevan, M., (2015). Analysis of the daily rainfall Events over India using a new long period (1901–2010) high resolution (0.25× 0.25) Gridded rainfall data set. Clim. Dyn. 45(3-4), 755–776

Pai, D.S., Sridhar, L., Rajeevan, M., Sreejith, O.P., Satbhai, N.S., Mukhopadhyay, B., (2014). Development of a new high spatial resolution (0.25× 0.25) long period (1901–2010) Daily gridded rainfall data set over India and its comparison with existing data sets Over the region. Mausam 65(1), 1–18.

Panda, R. K., Pramanik, N., and Bala, B. (2010). Simulation of river stage using artificial neural network and MIKE 11 hydrodynamic model. Computers & Geosciences, 36(6), 735-745.

Pandey, A., Himanshu, S. K., Mishra, S. K., & Singh, V. P. (2016). Physically based soil erosion and sediment yield models revisited. Catena, 147, 595-620.

Quilty, J., Adamowski, J., Khalil, B., and Rathinasamy, M. (2016). Bootstrap rank-ordered conditional mutual information (broCMI): A nonlinear input variable selection method for water resources modeling. *Water Resources Research*, *52*(3), 2299-2326.

Rajasekaran, S., Gayathri, S., and Lee, T. L. (2008). Support vector regression methodology for storm surge predictions. *Ocean Engineering*, *35*(16), 1578-1587.

Rathnayaka, K., Malano, H. and Arora, M., (2016). Assessment of sustainability of urban water supply and demand management options: a comprehensive approach. Water, 8(12), p.595.

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

Sapriza-Azuri, G., Jódar, J., Navarro, V., Slooten, L.J., Carrera, J. and Gupta, H.V., (2015). Impacts of rainfall spatial variability on hydrogeological response. *Water Resources Research*, *51*(2), pp.1300-1314.

Schalkoff, R. J. (1997). *Artificial neural networks*. McGraw-Hill Higher Education.

Schmid, M. D. (2009). A neural network package for Octave, User's Guide, Version: 0.1. 9.1.

Schwing, F. B., Murphree, T., and Green, P. M. (2002). The Northern Oscillation Index (NOI): a new climate index for the northeast Pacific. *Progress in Oceanography*, *53*(2-4), 115-139.

Sekhar, M., Shindekar, M., Tomer, S. K., and Goswami, P. (2013). Modeling the vulnerability of an urban groundwater system due to the combined impacts of climate change and management scenarios. *Earth Interactions*, *17*(10), 1-25.

Sekhar, M., Tomer, S., Thiyaku, S., Giriraj, P., Murthy, S., and Mehta, V. (2018). Groundwater Level Dynamics in Bengaluru City, India. *Sustainability*, *10*(1), 26.

745 Shiri, J., and KişI, Ö. (2011). Comparison of genetic programming with neuro-fuzzy systems
746 for predicting short-term water table depth fluctuations. *Computers & Geosciences*, *37*(10),
747 1692-1701.

748 Singh, K. P., Gupta, S., and Mohan, D. (2014). Evaluating influences of seasonal variations
749 and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning
750 approaches. *Journal of Hydrology*, *511*, 254-266.

751 Sivapragasam, C., Maheswaran, R., and Venkatesh, V. (2008). Genetic programming approach
752 for flood routing in natural channels. *Hydrological Processes: An International Journal*, *22*(5),
753 623-628.

754 Srivastava, A.K., Raajeevan, M., Kshirsagar, S.R., (2009). Development of a high resolution
755 Daily gridded temperature data set (1969–2005) for the Indian region. Atmos. Sci. Lett.
756 10(October), 249–254.

757 Sun, Y., Wendi, D., Kim, D.E., Liong, S.Y., (2016). Application of artificial neural networks
758 in groundwater table forecasting – a case study in a Singapore swamp forest. Hydrol. Earth
759 Syst. Sci. 20, 1405–1412

760 Suryanarayana, C. and Mahammood, V. (2019). Groundwater-level assessment and prediction
761 using realistic pumping and recharge rates for semi-arid coastal regions: a case study of
762 Visakhapatnam city, India. Hydrogeology journal, 27(1), 249-272.

763 Suryanarayana, C., Sudheer, C., Mahammood, V., and Panigrahi, B. K. (2014). An integrated
764 wavelet-support vector machine for groundwater level prediction in Visakhapatnam,
765 India. *Neurocomputing*, *145*, 324-335.

766 Tang, Q., Kurtz, W., Schilling, O.S., Brunner, P., Vereecken, H., and Franssen, H.J.H. (2017).
767 The influence of riverbed heterogeneity patterns on river-aquifer exchange fluxes under
768 different connection regimes. *Journal of hydrology*, *554*, pp.383-396.

769 Taylor, R.G., Scanlon, B., Döll, P., Rodell, M., Van Beek, R., Wada, Y., Longuevergne, L.,
770 Leblanc, M., Famiglietti, J.S., Edmunds, M., and Konikow, L. (2013). Ground water and
771 climate change. *Nature climate change*, *3*(4), p.322.

772 Todd, D.K., and Mays, L.W., (2005). Groundwater Hydrology, Third Revision John Wiley and
773 Sons Inc. 636p.

774 Vapnik, V., (1995). The nature of statistical learning theory. New York, NY: Springer-Verlag.

775 Vapnik, V., (2000). The nature of statistical learning theory. 2nd ed. New York, NY: Springer
776 Verlag.

777 Vapnik, V.N. and Vapnik, V., (1998). Statistical learning theory. New York, NY: Wiley, Vol.
778 1.

779 Vautard, R., and Ghil, M. (1989). Singular spectrum analysis in nonlinear dynamics, with
780 applications to paleoclimatic time series. *Physica D: Nonlinear Phenomena*, *35*(3), 395-424.

781 Velasco, E. M., Gurdak, J. J., Dickinson, J. E., Ferré, T. P. A., and Corona, C. R. (2017).
782 Interannual to multidecadal climate forcings on groundwater resources of the US West
783 Coast. *Journal of Hydrology: Regional Studies*, *11*, 250-265.

784 Vergara, J. R., and Estévez, P. A. (2014). A review of feature selection methods based on
785 mutual information. *Neural computing and applications*, *24*(1), 175-186.

786 Wang, S., Shao, J., Song, X., Zhang, Y., Huo, Z., and Zhou, X. (2008). Application of
787 MODFLOW and geographic information system to groundwater flow simulation in North
788 China Plain, China. *Environmental Geology*, *55*(7), 1449-1462.

789 Wang, W. C., Chau, K. W., Cheng, C. T., and Qiu, L. (2009). A comparison of performance
790 of several artificial intelligence methods for forecasting monthly discharge time series. *Journal*
791 *of Hydrology*, *374*(3-4), 294-306.

792 Wang, W. C., Chau, K. W., Xu, D. M., and Chen, X. Y. (2015). Improving forecasting accuracy
793 of annual runoff time series using ARIMA based on EEMD decomposition. *Water Resources*
794 *Management*, *29*(8), 2655-2675.

795 Wang, Y., Guo, S., Chen, H., and Zhou, Y. (2014). Comparative study of monthly inflow
796 prediction methods for the Three Gorges Reservoir. *Stochastic Environmental Research and*
797 *Risk Assessment*, *28*(3), 555-570.

798 Woodward, S.J., Wöhling, T. and Stenger, R., (2016). Uncertainty in the modelling of spatial
799 and temporal patterns of shallow groundwater flow paths: the role of geological and
800 hydrological site information. Journal of hydrology, 534, pp.680-694.

801 World Population Review (2018), Bengaluru Population Data (Urban Area) [Available at:
802 http://worldpopulationreview.com/world-cities/Bengaluru-population/]

803 Wu, C. L., and Chau, K. W. (2011). Rainfall–runoff modeling using artificial neural network
804 coupled with singular spectrum analysis. *Journal of Hydrology*, *399*(3-4), 394-409.

805 Wu, Chau, K. W., and Li, Y. S. (2009). Methods to improve neural network performance in
806 daily flows prediction. *Journal of Hydrology*, *372*(1-4), 80-93.

807 Wu, M. C., Lin, G. F., and Lin, H. Y. (2014). Improving the forecasts of extreme streamflow
808 by support vector regression with the data extracted by self-organizing map. *Hydrological*
809 *Processes*, *28*(2), 386-397.

810 Wunsch, A., Liesch, T., and Broda, S. (2018). Forecasting groundwater levels using nonlinear
811 autoregressive networks with exogenous input (NARX). *Journal of Hydrology*, *567*, 743-758.

812 Xu, T., Valocchi, A.J., Ye, M., and Liang, F., (2017). Quantifying model structural error:
813 Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a
814 data-driven error model. Water Resources Research, 53(5), pp.4084-4105.

815 Yadav, B., and Eliza, K. (2017). A hybrid wavelet-support vector machine model for prediction
816 of Lake water level fluctuations using hydro-meteorological data. *Measurement*, *103*, 294-301.

817 Yadav, B., and Mathur, S. (2018). River discharge simulation using variable parameter
818 McCarthy–Muskingum and wavelet-support vector machine methods. *Neural Computing and*
819 *Applications*, 1-14.

820 Yadav, B., Ch, S., Mathur, S., and Adamowski, J. (2016). Estimation of in-situ bioremediation
821 system cost using a hybrid Extreme Learning Machine (ELM)-particle swarm optimization
822 approach. *Journal of Hydrology*, *543*, 373-385.

823 Yadav, B., Ch, S., Mathur, S., and Adamowski, J., (2017). Assessing the suitability of extreme
824 learning machines (ELM) for groundwater level prediction. Journal of water and land
825 development, 32(1), pp.103-112.

826 Yang, C. T., Marsooli, R., and Aalami, M. T. (2009). Evaluation of total load sediment
827 transport formulas using ANN. *International Journal of Sediment Research*, *24*(3), 274-286.

828 Yao, X., Tham, L. G., and Dai, F. C. (2008). Landslide susceptibility mapping based on support
829 vector machine: a case study on natural slopes of Hong Kong, China. *Geomorphology*, *101*(4),
830 572-582.

831 Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O., and Lee, K. K. (2011). A comparative study of
832 artificial neural networks and support vector machines for predicting groundwater levels in a
833 coastal aquifer. *Journal of Hydrology*, *396*(1-2), 128-138.

834 Yousefi, H., Zahedi, S., Niksokhan, M. H., and Momeni, M. (2019). Ten-year prediction of
835 groundwater level in Karaj plain (Iran) using MODFLOW2005-NWT in
836 MATLAB. *Environmental Earth Sciences*, *78*(12), 343.

837 Zeng, Y., Xie, Z., and Zou, J. (2017). Hydrologic and climatic responses to global
838 anthropogenic groundwater extraction. *Journal of Climate*, *30*(1), pp.71-90.

839 Zhou, T., Wang, F., and Yang, Z. (2017). Comparative analysis of ANN and SVM models
840 combined with wavelet preprocess for groundwater depth prediction. *Water*, *9*(10), 781.