Environ Earth Sci (2015) 73:3119–3136
DOI 10.1007/s12665-014-3608-8

ORIGINAL ARTICLE

On the statistical forecasting of groundwater levels in unconfined aquifer systems

Sasmita Sahoo · Madan K. Jha

Received: 18 December 2013 / Accepted: 11 August 2014 / Published online: 16 September 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract The efficacy of multiple linear regression (MLR) as a modeling tool was investigated for forecasting groundwater levels in unconfined aquifer systems. The monthly groundwater level data of 17 sites and monthly data of rainfall, river stage and temperature and seasonal dummy variables were considered as input variables for MLR modeling. Three different approaches based upon plausible combinations of these input variables were considered to develop MLR models for individual sites. The regression coefficients of each MLR model following all the approaches were determined and the performance of these regression models was assessed using multiple correlation coefficient (R), multiple R^2, adjusted R^2, F-statistic, p-level, and standard error of estimate goodness-of-fit statistics. The best MLR models obtained for individual sites were then calibrated and validated to forecast groundwater levels over the study area. The analysis of the modeling results indicated that the MLR models developed in this study are able to predict groundwater levels with a reasonable accuracy at almost all the sites under study. It is concluded that the MLR technique can serve as a cost-effective and easy-to-use groundwater modeling tool for hydrogeologists, especially under inadequate field data condition.

Keywords Multiple linear regression · Groundwater forecasting · Statistical modeling · Stepwise regression · Unconfined aquifer

S. Sahoo (✉) · M. K. Jha
AgFE Department, Indian Institute of Technology Kharagpur, Kharagpur 721 302, West Bengal, India
e-mail: sasmitaiit@gmail.com

M. K. Jha
e-mail: madan@agfe.iitkgp.ernet.in

Introduction

A reliable, simple and practical approach for estimating groundwater level responses is of immense importance for proper planning and management of groundwater resources. Therefore, a variety of simulation and prediction techniques have been proposed for providing long-term measures to tackle the mismanagement and over-exploitation of groundwater resources. Because of the complex and dynamic nature of the groundwater flow phenomenon, it becomes very difficult to simulate groundwater level responses. Generally, several analytical and numerical techniques are available to simulate these complex groundwater flow processes. However, both the methods possess serious drawbacks when it comes to the practical assessments of groundwater resources. Analytical methods are limited in applicability to relatively simple hydrogeologic settings. Moreover, it involves formulation of differential equations and extensive calculations which require substantial computational resources to generate multiple scenarios. Physically-based numerical models of groundwater flow are capable of simulating complex hydrogeological settings and have the potential for detailed simulation of groundwater levels using hydrologic variables, but require huge amount of good quality information to understand the basic physical process and to characterize the hydrologic system, much of which is unknown, uncertain, or difficult to obtain. Furthermore, the development of numerical models involves sophisticated programs for calibration using rigorous optimization techniques, which is time consuming, costly, and not certain to produce reasonable results. Therefore, in such cases, empirically-based models are proven to be effective in modeling groundwater fluctuations as they use simple equations to model the relationship between the parameters and do not require the mathematical formulation of the complex underlying phenomenon. Multiple linear regression (MLR) models are such models which can simulate the association between the explanatory and response variables by fitting a linear equation to the observed data (Makridakis et al. 2008) and can provide useful results using relatively less data, and are less laborious and cost-effective. Moreover, it has the flexibility in inclusion of any number of independent variables in the model. Despite the incapability of MLR models to deal with the non-linearity existing between model inputs and outputs, they have been used in many hydrological studies probably due to the fact that the results are quite easy to use and also the interpretation of the relationship between the parameters is easier (Heuvelmans et al. 2006; Adeloye 2009).

Most studies reported in the literature rely on linear regression techniques to establish the relationship, although it is well known that the underlying assumptions, e.g., the linearity of the relationship between the dependent variable and model parameters, are often violated. Until the early 1990s, the MLR technique was used by NRCS (Natural Resources Conservation Service, USA) for many years for streamflow forecasting. In the recent past, some researchers have used advanced statistical techniques such as principal component regression (PCR) for streamflow forecasting (e.g., Eldaw et al. 2003; Hsieh et al. 2003; Risley et al. 2005; Tootle et al. 2007). Heuvelmans et al. (2006) compared linear regression and ANN techniques for regionalizing the most sensitive parameters of the semi-distributed hydrological model SWAT (Soil and Water Assessment Tool). A few applications of MLR technique in hydrological modeling have been reported for the prediction of sediment load concentration (Sinnakaudan et al. 2006), prediction of surface water storage-yield-reliability relationship (Adeloye 2009), forecasting of water supply (e.g., McCuen et al. 1979; USDA-NRCS 2007) and the prediction of reference evapotranspiration (Xu et al. 2012). However, the use of statistical techniques in groundwater modeling is very limited. Hodgson (1978) used multiple linear regression (MLR) as a modeling tool for the simulation of water table responses in the Vryburg aquifer of South Africa considering rainfall and pumping as input parameters. Shao and Campbell (2002) used regression methods to model trends in groundwater levels. Uddameri (2007) used regression and artificial neural network (ANN) techniques to forecast piezometric levels in a deep well of South Texas using bias, time and dummy variables as inputs to the models; no real-world data have been used for developing ANN and MLR models in this study. Sahoo and Jha (2013) examined the potential of two data-driven approaches, MLR and ANN, for simulating transient groundwater levels in an unconfined aquifer system. On the other hand, some researchers have also applied regression techniques in the field of groundwater quality modeling. For instance, Ahsan et al. (2008) used MLR to model groundwater arsenic contamination in Bangladesh. Nolan and Hitt (2006), Stigter et al. (2008) and Boy-Roura et al. (2013) applied nonlinear regression technique, factorial regression technique and MLR technique, respectively, for predicting nitrate contamination in aquifers. Chen and Druliner (1988), Steichen et al. (1988), Stackelberg et al. (2006, 2012) used different types of regression models including MLR to predict pesticide concentrations in groundwater. Some more applications of regression techniques in the fields of environmental earth science and hydrology have been recently reported by some researchers (e.g., Abdelgawad et al. 2010; Cui and Du 2011; Bayramov et al. 2012; Kayabasi 2012; Liu et al. 2013 and Siddiqui and Osman 2013).

From the above review and literature, it is evident that very limited study has been carried out till date on the prediction of groundwater levels using multiple linear regression (MLR) technique. In addition, to the best knowledge of the authors, no methodology exists in the literature for finding out significant influential inputs to develop MLR models using real-world data. Therefore, the motivation for this research is two-fold. Firstly, to identify the potential/influential inputs (hydrologic variables) for statistical forecasting of groundwater levels in an unconfined aquifer system. Secondly, to develop site-specific MLR models for forecasting groundwater levels using the influential hydrologic variables. To accomplish these objectives effectively, the inputs to the MLR models are selected considering three approaches based upon different combinations of the influential input variables. The effectiveness of MLR models developed using three approaches are evaluated using standard statistical indicators. Finally, the best MLR models are calibrated and validated to forecast groundwater levels at individual sites over the study area. Thus, this study demonstrates a scientifically-sound methodology for the development and evaluation of MLR models in forecasting groundwater levels using real-world data.

Methodology

Collection of Hydrological and Hydrogeological data

The study area is a part of Konan groundwater basin located in Kochi Prefecture of Shikoku Island, Japan (Fig. 1) extending over an area of 2,200 ha. Of the total area, about 1,502.5 ha is paddy fields, 488.0 ha upland, and 186.5 ha is under greenhouse cultivation. It is bounded by the Monobe River (perennial) in the west and the Koso River (intermittent) in the east. The average daily

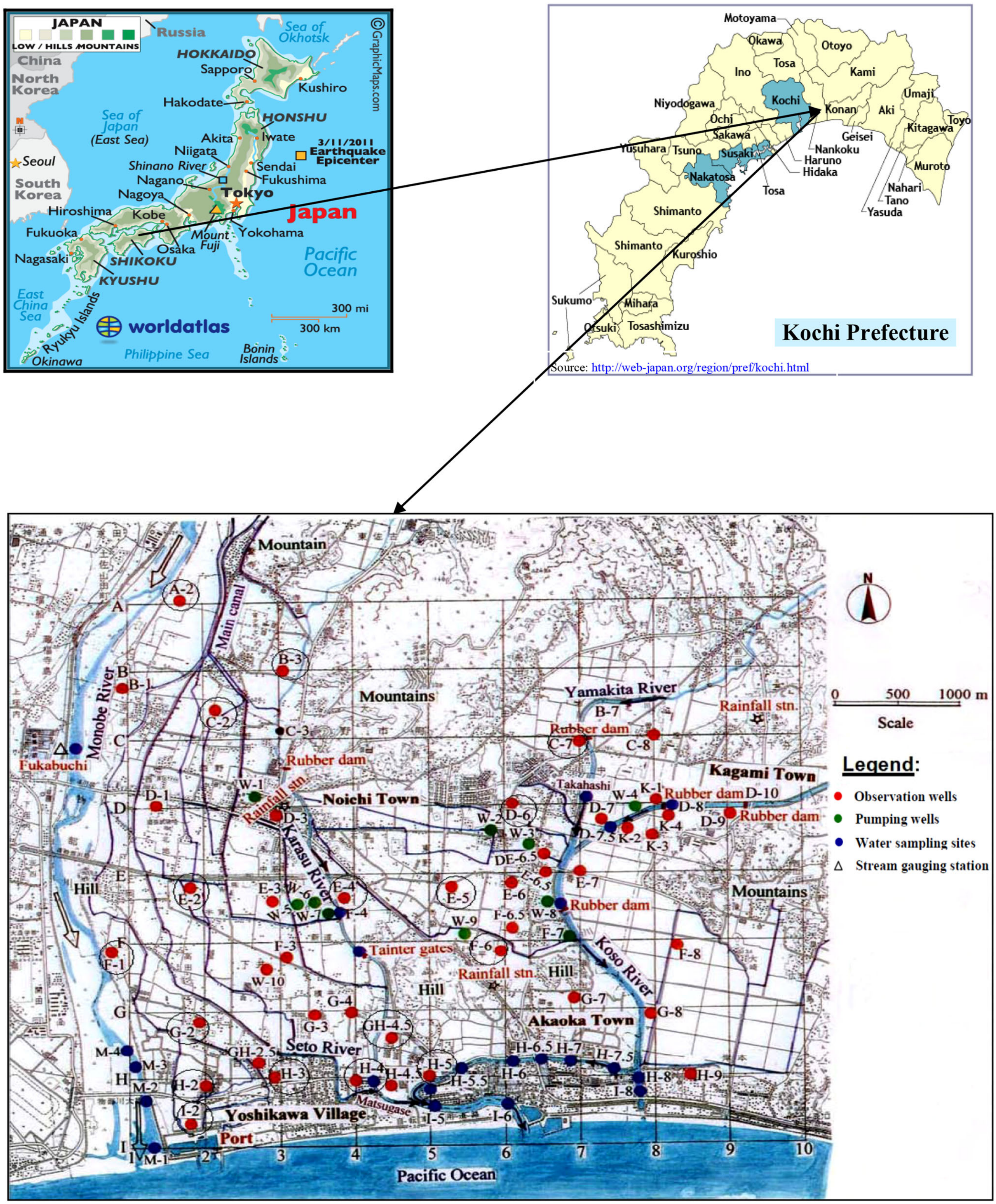

Fig. 1 Map of the Konan groundwater basin with the location of observed sites

maximum temperature is 37 °C during summer, and the average daily minimum temperature is −4 °C during winter. The mean annual rainfall and evapotranspiration in the study area are about 2,600 mm and 800 mm, respectively. The overall flow of groundwater in the basin is from north to south into the Pacific Ocean (Jha et al. 1999), with significant Monobe River-aquifer interaction up to a larger portion of the basin (Jha et al. 2004).

A total of 17 sites (A-2, B-3, C-2, C-7, D-6, E-2, E-4, E-5, F-1, F-6, G-2, GH-4.5, H-2, H-3, H-4, H-5 and I-2) were selected over the study area on the basis of the availability and continuity of long-term groundwater-level data at a site. The location of these sites is shown in Fig. 1 as encircled observation wells. It is clear from Fig. 1 that the selected sites are representative of the unconfined portion of the study area. For carrying out MLR modeling, the monthly data of rainfall, temperature, river stage and groundwater level for 6 years (1999–2004) were considered. The monthly groundwater-level data for 17 selected sites were obtained for the 1999–2004 period from the Kochi Prefectural Office, Kochi City, Japan. The monthly rainfall data and monthly temperature data of 6 years (1999–2004) were gathered from Nankoku-shi, Japan. The monthly river-stage data of the Monobe River at the Fukabuchi gauging station (Fig. 1) for the 1999–2004 period were collected from the Kochi Work Office, the Ministry of Construction, Japan. Figure 2a, b and c illustrates the monthly variation in the rainfall, temperature and river stage during the study period (1999–2004), together with the bar graphs of standard errors. The standard error bars indicate the temporal variation of rainfall, temperature and river stage in a particular month over the 1999–2004 period.

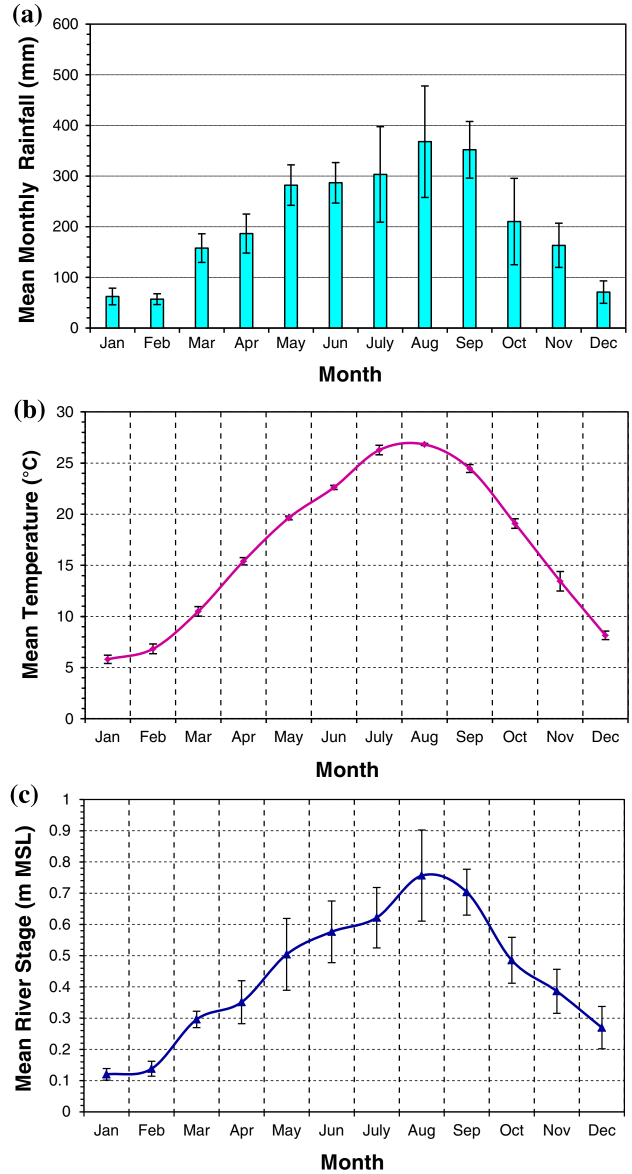

Fig. 2 **a** Mean monthly rainfall in the study area during 1999–2004 with standard *error bars*. **b** Mean monthly temperature in the study area during 1999–2004 with standard *error bars*. **c** Mean monthly river stage in the study area during 1999–2004 with standard *error bars*

Conceptual model of the study area for MLR modeling

A key step in the modeling procedure is to develop or build a conceptual model of the system to be modeled. The purpose of building a conceptual model is to simplify the complex field problem (complex natural process) and make it more amenable to modeling. The study area is predominated by an unconfined aquifer which consists of alluvial sand and gravel. The hydraulic conductivity of this aquifer varies from 65 to 804 m/day (Jha et al. 1999). The environmental factors affecting groundwater levels in an unconfined aquifer could be numerous (Todd 1980), of which rainfall, river stage, evapotranspiration, pumping, tides are important factors which are expected to influence the unconfined aquifer-groundwater level of the study area. However, because of the unavailability of some of the data (environmental factors affecting groundwater level) in the study area, rainfall, river stage and temperature (often considered as a surrogate of evapotranspiration) were used as model inputs together with seasonal dummy variables and significant lags of rainfall, temperature, river stage and groundwater level. In this study, multiple linear regression (MLR) models have been developed for predicting/forecasting groundwater levels. A schematic of the conceptual model of the study area (unconfined aquifer under study) for MLR modeling is shown in Fig. 3. The development of empirical models like the MLR model requires the identification of suitable hydrologic/hydrogeologic factors (time series of inputs) that have significant influence on groundwater and the outputs (time series of observed groundwater levels) as illustrated in Fig. 3.

Multiple linear regression technique

Multiple linear regression (MLR) technique establishes the relationship between two or more explanatory

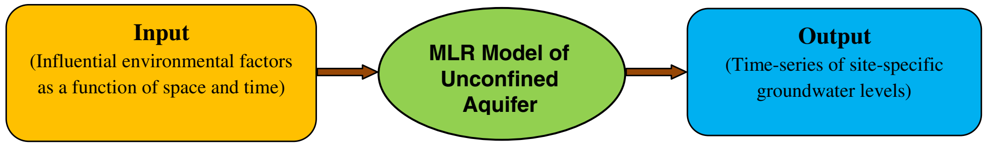

Fig. 3 Schematic representation of the MLR model used in the study

(independent) variables and a response (dependent) variable by fitting a linear equation to the observed data. The general form of a MLR model is given as (Makridakis et al. 2008):

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots \cdots\cdots + \beta_k X_{k,i} + \varepsilon_i \quad (1)$$

where Y_i represent the ith observations of each of the dependent variable Y; $X_{1,i}$, $X_{2,i}$ …, $X_{k,i}$ represent the ith observations of each of the independent variables X_1, X_2, …, X_k respectively; β_0, β_1, β_2, …, β_n are fixed (but unknown) parameters; and ε_i is a random variable that is normally distributed.

The task of regression modeling is to estimate the unknown parameters (β_0, β_1, β_2, …, β_n) of the MLR model [Eq. (1)]. Thus, the pragmatic form of the statistical regression model obtained after applying the least square method is as follows (Makridakis et al. 2008).

$$Y_i = b_0 + b_1 X_{1,i} + b_2 X_{2.i} + \cdots \cdots\cdots + b_k X_{k,i} + e_i \quad (2)$$

$$\text{Therefore, estimate of } Y = \hat{Y} = b_0 + b_1 X_{1,i} + b_2 X_{2.i} + \cdots \cdots\cdots + b_k X_{k,i} \quad (3)$$

where, $i = 1, 2, \ldots, n$; b_0, b_1, b_2,… b_k, are the estimates or unstandardized regression coefficients of β_0, β_1, β_2, …, β_n respectively; and e_i is the estimated error (or residual) for the ith observation. Further details of MLR technique can be found in standard literature (Makridakis et al. 2008; Spiegel and Stephens 2008).

Identification of input variables for MLR models

In this study, MLR models were developed for predicting groundwater levels at 17 sites considering three approaches; these approaches are based on different sets of inputs used for developing MLR models (Table 1). In Approach I, only hydrological variables (rainfall, temperature and river stage) are used as inputs for developing MLR models which are easily obtainable. These inputs (Approach I) were decided by cross correlation analysis and simple linear regression analyses between monthly groundwater levels at individual sites and monthly cumulative rainfalls, monthly river stage and monthly mean temperature, respectively. The cross correlation and linear regression analyses indicated that the rainfall, temperature and river stage have a good correlation with the groundwater levels of all the 17 sites (Jha and Sahoo 2014).

Table 1 Sets of inputs considered for developing MLR models

Approach	Inputs considered	Total no. of inputs
I	Hydrological variables (Rainfall, temperature, and river stage) for 17 sites	3
II	Hydrological variables and seasonal dummies (D_1–D_{11}) for 17 sites	14
III	Hydrological variables, seasonal dummies, and 1-month and 2-month lags of hydrological variables and groundwater level for 17 sites	22

Here, it is worth mentioning that the positive correlation between monthly temperature and monthly groundwater levels obtained in this study is due to the fact that the months having high or low temperature coincide with the months having high or low rainfall and river stage (Fig. 2a, b, c). The less temporal variability of monthly groundwater level and monthly temperature in the study area could be the possible reason for better correlation between these two parameters. Temperature is often considered as an input to an empirical model of an unconfined aquifer in order to indirectly take into account the influence of evapotranspiration on groundwater level because evapotranspiration is difficult to measure on a large scale and hence not easily available in most cases. Thus, considering temperature as a surrogative to ET in absence of field measured values of loss of groundwater by ET, is a widely accepted practice in the area of empirical modeling of groundwater levels (e.g., Coulibaly et al. 2001; Coppola et al. 2003, 2005; Daliakopoulos et al. 2005; Nourani et al. 2008; Trichakis et al. 2009). Furthermore, empirical models being a black box model do not involve water balance computation and therefore, the nature of influence (positive/negative) of the environmental factors (natural and anthropogenic) on the groundwater level is not considered by such models.

In Approach II, apart from the hydrological variables, seasonal dummy variables were also considered as inputs to the MLR models. The dummy variables are a set of time-related explanatory variables (i.e., seasonal dummy variables) which were introduced in the regression models to investigate the effect of seasonal fluctuation on the groundwater levels over the study area. To handle the seasonality in the monthly dataset, 11 (i.e., 12 − 1) dummy variables were considered to denote 12 months. Each of the 11 seasonal dummy variables indicates one new independent variable having only two allowable values, 0 or 1, which can be defined as follows (Makridakis et al. 2008):

$$D_1 = \begin{cases} 1, \text{ if the month is January} \\ 0, \text{ if otherwise} \end{cases}; D_2 = \begin{cases} 1, \text{ if the month is February} \\ 0, \text{ if otherwise} \end{cases}; D_3 = \begin{cases} 1, \text{ if the month is March} \\ 0, \text{ if otherwise} \end{cases};$$

$$\text{and so on; with } D_{11} = \begin{cases} 1, \text{ if the month is November} \\ 0, \text{ if otherwise} \end{cases}$$

The regression coefficients associated with seasonal dummy variables is a measure of the difference in the groundwater level between those months and the omitted month (i.e., December), which is chosen as the base period. For example, the coefficient of D_1 reflects the effect of (January–February) change in the groundwater level at a particular site compared to December. Similarly, D_2 rep-

Table 2 Statistical indicators used to evaluate the efficacy of the developed MLR models

Sl. no.	Statistical indicators	Significance	Mathematical expression
1	Standardized regression coefficient (β_j)	The magnitude of 'β' indicates the relative contribution of each independent variable in the prediction of the dependent variable	$\beta_j = b_j \frac{\sigma_j}{\sigma_Y}$
		It measures the impact of a unit change in the standardized value of independent variable on the standardized value of dependent variable	
2	Standard error (s.e.)	It is a measure of the stability of regression coefficients (Spiegel and Stephens 2008)	$\text{S.E.} = \frac{\sigma}{\sqrt{n-1}}$
3.	t test	The t test is done for each regression coefficient to examine its significance in the presence of all other explanatory variables (Makridakis et al. 2008)	$t = \frac{bj}{\text{S.E.}(bj)}$
		The higher the calculated t value, the stronger is the significance of the regression coefficient in a model	
4	F test	The overall statistical significance of the regression model is evaluated using an F test in the format of analysis of variance (Makridakis et al. 2008)	$F = \frac{(\text{SSR}/p)}{(\text{SSE}/(n-p-1))} = \frac{\text{MSR}}{\text{MSE}}$
5	Coefficient of multiple determination (R^2)	The fit of a MLR model is assessed by R^2, which represents the proportion of total variation of the dependent variable accounted for or, explained by the independent variables	$R^2 = \frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2} = \frac{\text{Explainedsumofsquares}}{\text{Totalsumofsquares}} = \frac{\text{SSR}}{\text{SST}}$
6	Multiple correlation coefficient (R)	It is the correlation between the observed value of a dependent variable (Y) and an estimated or predicted value of Y based on multiple explanatory variables	Squared root of R^2
7	Adjusted R^2	It is interpreted similarly to the R^2 value except the 'adjusted $R^{2\prime}$ takes into consideration the number of degrees of freedom (Statistica 2001)	Adjusted $R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2)$
		The model with highest value of adjusted R^2 is considered as the best regression model	
8	p-level	The p-level represents the probability of error that is involved in accepting the observed result as valid	–
		The lower the p-level, the stronger is the statistical significance	
9	Standard error of estimate (SEE)	SEE is the measure of the amount of error in the prediction of dependent variable (Y) for each independent variable (X) in the regression equation.	$\text{SEE} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-p-1}}$

b_j unstandardized regression coefficient, σ_j standard deviations associated with the jth independent variable, σ_Y standard deviations associated with the dependent variable Y, σ sample standard deviation, n number of observations, p number of independent variables, *SSR* sum of squares regression, *SSE* sum of squared error, *SST* total sum of squares, *MSR* mean squared regression, *MSR* mean squared error, Y_i observed value of a dependent variable, $\hat{Y}_i$ estimated or predicted value of Y, $\bar{Y}$ mean value of a dependent variable

Table 3 MLR coefficients and their significance for the MLR model of Site B-3 using standard and stepwise regression techniques

Parameters	Beta (β)	SE (m)	b	SE (m)	t (69)	p-level
Standard regression						
Intercept			19.6227	0.017	1,098.76	0.0000
Rainfall	0.589	0.127	0.0004	0.0001	4.61	0.0000
Temperature	0.269	0.104	0.0039	0.002	2.59	0.0117
River stage	0.040	0.147	0.0170	0.060	0.27	0.7835
Regression equation: GWL $= 19.6227 + 0.0004 \times R + 0.0039 \times T + 0.0170 \times S$						
Stepwise regression						
Intercept			19.6229	0.0170	1,107.71	0.0000
Rainfall	0.614	0.089	0.0004	0.0001	6.87	0.0000
Temperature	0.284	0.089	0.0041	0.0013	3.17	0.0022
Regression equation: GWL $= 19.6229 + 0.0004 \times R + 0.0041 \times T$						

β Standardized regression coefficient, *SE* Standard error, *b* Estimates or unstandardized regression coefficient, *GWL* Groundwater level, *R* Rainfall, *T* Temperature, *S* River stage

Table 4 ANOVA table for the MLR model of Site B-3 using standard and stepwise regression techniques

Sources of variation	Sums of squares	Degree of freedom	Mean squares	F-statistic	p-level
Standard regression					
Regression	0.561	3	0.187	49.82	0.0000
Residual	0.255	68	0.004		
Total	0.817				
Stepwise regression					
Regression	0.561	2	0.280	75.69	0.0000
Residual	0.255	69	0.004		
Total	0.816				

Table 5 Goodness-of-fit statistics of the MLR model for Site B-3 using standard and stepwise regression technique

Statistics	Standard regression	Stepwise regression
Multiple R	0.8270	0.8288
Multiple R^2	0.6839	0.6869
Adjusted R^2	0.6735	0.6779
F	49.82	75.69
p	0.0000	0.0000
SEE (m)	0.0613	0.0609

R Correlation coefficient, R^2 Coefficient of determination, *SEE* Standard error of estimate

Table 6 Regression coefficients of MLR models for predicting groundwater levels at 17 sites using Approach I (stepwise regression technique)

Site	Regression coefficients			
	Intercept	Rainfall	Temperature	River stage
A-2	16.1635	0.0004	0.0255	0.6758
B-3	19.6229	0.0004	0.0041	–
C-2	12.0699	0.0010	0.1408	–
C-7	8.0665	–	0.0040	−0.1980
D-6	5.0821	0.0004	0.0473	0.4444
E-2	5.0191	–	0.0991	0.7461
E-4	4.4757	0.0009	0.1286	–
E-5	3.8284	0.0008	0.1263	–
F-1	2.3299	0.0003	0.0252	0.9139
F-6	3.0275	–	0.0787	0.4031
G-2	0.7682	0.0004	0.0872	0.5862
GH-4.5	−0.2248	–	0.0697	0.8184
H-2	0.1279	–	0.0307	0.3575
H-3	0.2304	–	0.0739	0.56751
H-4	0.3493	–	0.0588	0.3905
H-5	−0.3394	–	0.5615	0.0190
I-2	−0.0052	–	0.0174	0.2506

resents the effect of (February–March) change in the groundwater level at a site, D_3 refers to the effect of (March–April) change in the groundwater level, and so on, with D_{11} indicating the effect of (November–December) change in the groundwater level.

In Approach III, all the input variables considered in Approach II and significant lags of rainfall, temperature, river stage and groundwater level were considered as inputs to the MLR models. The impact of time lags of hydrological and hydrogeological variables (rainfall, temperature, and river stage and groundwater level) has a greater importance on the prediction of groundwater level. Therefore, the influence of multi-period lags of rainfall, temperature and river stage on groundwater levels at 17 sites was examined by cross-correlation technique. Similarly, the effect of previous month's groundwater levels on the current month's groundwater levels was analyzed by partial autocorrelation techniques with the help of

Table 7 Goodness-of-fit statistics of MLR models developed for 17 sites using Approach I

Sites	Multiple R	Multiple R^2	Adjusted R^2	F-statistic	p-level	SEE (m)
A-2	0.956	0.915	0.911	242.85	0.0000	0.125
B-3	0.829	0.687	0.678	75.70	0.0000	0.060
C-2	0.887	0.786	0.780	126.66	0.0000	0.615
C-7	0.400	0.160	0.135	6.56	0.0025	0.082
D-6	0.902	0.813	0.805	98.50	0.0000	0.246
E-2	0.898	0.806	0.800	143.23	0.0000	0.445
E-4	0.922	0.850	0.846	195.87	0.0000	0.451
E-5	0.941	0.885	0.882	266.03	0.0000	0.375
F-1	0.957	0.917	0.913	248.95	0.0000	0.140
F-6	0.946	0.895	0.892	293.66	0.0000	0.233
G-2	0.936	0.875	0.870	158.94	0.0000	0.319
GH-4.5	0.787	0.619	0.608	56.03	0.0000	0.553
H-2	0.892	0.795	0.789	134.00	0.0000	0.158
H-3	0.912	0.832	0.827	170.61	0.0000	0.306
H-4	0.861	0.742	0.734	99.03	0.0000	0.310
H-5	0.736	0.542	0.529	40.82	0.0000	0.253
I-2	0.892	0.795	0.790	134.11	0.0000	0.094

R Correlation coefficient, R^2 Coefficient of determination, *SEE* Standard error of estimate

STATISTICA software. Based on the results of cross-correlation and partial autocorrelation, significant lags were identified. Using the significant lags of hydrological and hydrogeological variables, influential hydrological variables and 11 seasonal dummy variables, MLR models were developed for each of the 17 sites (Approach III).

Development of MLR models

The 'Multiple Regression' tool of STATISTICA software was used to determine the coefficients of each of the 17 MLR models using three different approaches. Using these coefficients, the monthly groundwater levels for individual sites were forecasted. In this study, as a preliminary test, both standard and stepwise regression methods were used in Approach I to develop MLR models. Finally, stepwise regression technique was employed to develop MLR models for all the three approaches as it considers only most significant input variables for predicting groundwater levels. Then, the best among the three approaches were determined and 17 MLR models were developed using partial time series data considering the inputs of the best MLR models. Total available data were divided into two sets for calibration and validation of the best MLR models obtained among the above three approaches. The monthly groundwater level data of 4 years (1999–2002) were used for calibrating the MLR models and those of 2 years (2003–2004) for validating the models.

Evaluation of MLR models

The effectiveness of the MLR models was measured by a set of statistical indicators, viz., standardized regression coefficient (β), standard error, F test, t test, coefficient of multiple determination (R^2), multiple correlation coefficient (R), adjusted R^2, p-level and standard error of estimate (SEE). These indicators are briefly summarized in Table 2.

Apart from the quantitative evaluation, a visual comparison of the observed and estimated groundwater levels from the MLR models based on the three approaches was also performed using scatter plots with 1:1 line as well as by simple plots of observed and calculated groundwater levels.

Diagnostic checks for developed MLR models

After developing the MLR models following three approaches, they were checked for the basic assumptions to justify the use of multiple linear regression models for prediction. The four assumptions of MLR models are (Makridakis et al. 2008): (a) linearity of the relationship between dependent and independent variables (b) independence of the errors (no autocorrelation) (c) homoscedasticity (constant variance) of the errors with respect to the predicted values, and (d) normality of the error distribution.

Results and discussion

MLR models based on hydrological variables (Approach I)

A comparison of the results of multiple linear regression analysis using both standard regression and stepwise regression technique for the Site B-3 using rainfall, temperature and river stage are summarized in Tables 3, 4 and 5 as an example. It is obvious from Table 3 that in stepwise regression, the river stage has been removed from the regression equation due to its lower t-value and higher p-level, which indicates that it is a statistically insignificant variable for the prediction of groundwater level at Site B-3. Moreover, it is clear from Table 5 that adjusted R^2 value from stepwise regression is higher than that of standard

Table 8 MLR models for predicting groundwater levels at 17 sites using Approach II

Site	MLR models (Approach II)
A-2	$GWL = 16.1241 + 0.0003 \times R + 0.0255 \times T + 0.7686 \times S + 0.2257 \times D_4 + 0.0707 \times D_5 - 0.1045 \times D_9$
B-3	$GWL = 19.6150 + 0.0004 \times R + 0.0042 \times T + 0.0316 \times D_3 + 0.0384 \times D_4 + 0.0436 \times D_8 - 0.03142 \times D_9$
C-2	$GWL = 12.2676 + 0.0012 \times R + 0.1332 \times T - 0.5653 \times D_1 - 0.5891 \times D_2 - 0.3336 \times D_3 + 0.9949 \times D_4 + 0.3680 \times D_5 - 0.4904 \times D_8 - 0.6965 \times D_9$
C-7	$GWL = 8.0475 - 0.0001 \times R + 0.0046 \times T - 0.1192 \times S + 0.0625 \times D_4 - 0.0545 \times D_8$
D-6	$GWL = 5.0832 + 0.0004 \times R + 0.0475 \times T + 0.5747 \times S - 0.2662 \times D_2 - 0.1130 \times D_3 + 0.3037 \times D_4 - 0.2842 \times D_8 - 0.2521 \times D_9$
E-2	$GWL = 5.5349 + 0.0005 \times R + 0.0681 \times T + 0.5919 \times S - 0.5519 \times D_1 - 0.9117 \times D_2 - 0.3029 \times D_3 + 0.8014 \times D_4 + 0.3053 \times D_5 + 0.3036 \times D_7 - 0.2010 \times D_9$
E-4	$GWL = 5.0657 + 0.0011 \times R + 0.0967 \times T + \times S - 0.5265 \times D_1 - 0.9977 \times D_2 - 0.5106 \times D_3 + 0.6924 \times D_4 + 0.1525 \times D_5 + \times D_6 + 0.3141 \times D_7 - 0.2131 \times D_9$
E-5	$GWL = 4.4064 + 0.0006 \times R + 0.0845 \times T + 0.4088 \times S - 0.3261 \times D_1 + 0.6930 \times D_3 + 0.0845 \times D_4 + 0.1509 \times D_5 + 0.6528 \times D_7 + 0.2269 \times D_8 - 0.6058 \times D_9 - 1222 \times D_{11}$
F-1	$GWL = 2.3046 + 0.0003 \times R + 0.0258 \times T + 0.9913 \times S - 0.0811 \times D_2 + 0.2014 \times D_4 - 0.0876 \times D_8 + 0.1152 \times D_9$
F-6	$GWL = 3.4025 + 0.0003 \times R + 0.0573 \times T + 0.3112 \times S - 0.2117 \times D_1 - 0.5040 \times D_2 - 0.4286 \times D_3 + 0.3047 \times D_4 + 0.2752 \times D_7 + 0.0797 \times D_8$
G-2	$GWL = 1.3011 + 0.0007 \times R + 0.0557 \times T + 0.4834 \times S - 0.4483 \times D_1 - 0.7189 \times D_2 - 0.5062 \times D_3 + 0.4465 \times D_4 + 0.1052 \times D_6 + 0.3185 \times D_7 + 0.1662 \times D_8 + 0.1836 \times D_{10} + 0.1563 \times D_{11}$
GH-4.5	$GWL = 0.6537 + 0.0010 \times R + 0.0193 \times T - 0.9933 \times D_2 - 0.6262 \times D_3 + 0.7634 \times D_8 + 0.8433 \times D_9 + 0.9238 \times D_{10} + 0.8651 \times D_{11}$
H-2	$GWL = 0.3097 + 0.0244 \times T + 0.2148 \times S - 0.1658 \times D_1 - 0.2583 \times D_2 - 0.2471 \times D_3 - 0.1593 \times D_5 - 0.0877 \times D_6 + 0.1097 \times D_{10} + 0.1116 \times D_{11}$
H-3	$GWL = 0.7641 + 0.0006 \times R + 0.0540 \times T + 0.1840 \times S - 0.4777 \times D_1 - 0.7337 \times D_2 - 0.6917 \times D_3 - 0.2849 \times D_5 - 0.1556 \times D_6 + 0.2061 \times D_{10} + 0.2552 \times D_{11}$
H-4	$GWL = 0.8050 + 0.0006 \times R + 0.0448 \times T - 0.4470 \times D_1 - 0.6845 \times D_2 - 0.5852 \times D_3 - 0.2539 \times D_5 - 0.2393 \times D_6 - 0.1388 \times D_7 + 0.2609 \times D_{11}$
H-5	$GWL = -0.0397 + 0.0047 \times T + 0.3962 \times S - 0.2149 \times D_1 - 0.3566 \times D_2 - 0.2871 \times D_3 - 0.1302 \times D_4 - 0.1995 \times D_5 + 0.2445 \times D_8 + 0.3450 \times D_9 + 0.3903 \times D_{10} + 0.3266 \times D_{11}$
I-2	$GWL = 0.0877 + 0.0129 \times T + 0.2164 \times S - 0.0653 \times D_1 - 0.1092 \times D_2 - 0.1084 \times D_3 - 0.0711 \times D_5 + 0.0675 \times D_8 + 0.0745 \times D_9 + 0.1105 \times D_{10} + 0.0660 \times D_{11}$

regression, whereas SEE is less than that of standard regression; their significance is also supported by corresponding F-statistics (Table 4) and p-level. Since stepwise regression is giving better results as compared to the standard regression, all the analyses hereafter were performed using stepwise regression technique.

The regression coefficients of 17 MLR models determined by stepwise regression technique considering Approach I for predicting groundwater levels at 17 sites are presented in Table 6. It is clear from Table 6 that all the three hydrological variables (rainfall, temperature and river stage) are responsible for groundwater fluctuations at Sites A-2, D-6, F-1 and G-2, whereas the groundwater level is affected by only rainfall and temperature at Sites B-3, C-2, E-4 and E-5. On the other hand, groundwater levels of remaining sites (C-7, E-2, F-6, GH-4.5, H-2, H-3, H-4, H-5 and I-2) are affected by temperature and river stage. The goodness-of-fit statistics of stepwise regression for 17 sites by considering rainfall, temperature and river stage as input variables are summarized in Table 7. Here, Site F-1 shows highest multiple R and adjusted R^2 value and their significance is also supported by corresponding F-statistics and p-level, whereas Site C-7 has the lowest multiple R and adjusted R^2 among 17 sites. This indicates that the groundwater level at Site C-7 is not influenced by these factors and there may be some other factors which are responsible for the groundwater fluctuation.

MLR models using hydrological and dummy variables (Approach II)

MLR modeling was carried out for all the 17 sites by considering 14 independent variables (3 hydrological variables and 11 seasonal dummy variables), which are presented in Table 8. A comparison between the goodness-

Table 9 Comparison of MLR models with and without seasonal dummies

Site	MLR model with seasonal dummies					MLR model without seasonal dummies				
	Adj. R^2	MSE (m^2)	SEE (m)	*F*-statistic	*p*-level	Adj. R^2	MSE (m^2)	SEE (m)	*F*-statistic	*p*-level
A-2	0.939	0.011	0.103	183.14	0.0000	0.911	0.016	0.125	242.85	0.0000
B-3	0.697	0.004	0.059	28.16	0.0000	0.678	0.004	0.061	75.70	0.0000
C-2	0.894	0.183	0.428	67.22	0.0000	0.780	0.379	0.615	126.66	0.0000
C-7	0.178	0.006	0.080	4.08	0.0030	0.135	0.007	0.082	6.56	0.0030
D-6	0.875	0.039	0.196	63.21	0.0000	0.805	0.060	0.246	98.50	0.0000
E-2	0.933	0.067	0.259	99.11	0.0000	0.800	0.198	0.445	143.23	0.0000
E-4	0.938	0.082	0.286	120.35	0.0000	0.846	0.204	0.451	195.87	0.0000
E-5	0.952	0.057	0.239	118.11	0.0000	0.882	0.141	0.375	266.03	0.0000
F-1	0.934	0.015	0.122	144.67	0.0000	0.913	0.020	0.140	248.95	0.0000
F-6	0.954	0.023	0.152	163.06	0.0000	0.892	0.054	0.233	293.66	0.0000
G-2	0.949	0.040	0.199	110.58	0.0000	0.870	0.102	0.319	158.95	0.0000
GH-4.5	0.854	0.114	0.338	42.43	0.0000	0.608	0.306	0.553	56.03	0.0000
H-2	0.873	0.015	0.122	49.73	0.0000	0.789	0.025	0.158	134.00	0.0000
H-3	0.953	0.026	0.160	144.28	0.0000	0.827	0.093	0.306	170.61	0.0000
H-4	0.883	0.043	0.206	60.33	0.0000	0.734	0.097	0.310	99.03	0.0000
H-5	0.834	0.023	0.150	33.33	0.0000	0.529	0.064	0.253	40.82	0.0000
I-2	0.854	0.006	0.079	42.38	0.0000	0.790	0.009	0.094	134.11	0.0000

R^2 Coefficient of determination, *MSE* Mean squared error, *SEE* Standard error of estimate

of-fit statistics for the MLR models with seasonal dummies (Approach II) and those for the MLR models without seasonal dummies (Approach I), is illustrated in Table 9 to study the influence of seasonal dummy variables at 17 sites. It can be observed that the models with seasonal dummies possess improved values of goodness-of-fit statistics (higher adjusted R^2 and lower SEE) as compared to those without seasonal dummies for all the sites. Thus, it suggests that the seasonal dummy variables have significant influence on groundwater fluctuations at all the sites over the study area.

MLR models using hydrological variables, dummy variables and multi-period lags of hydrological and hydrogeological variables (Approach III)

The MLR models developed using the significant lags of hydrological and hydrogeological variables, influential hydrological variables and 11 seasonal dummy variables for each of the 17 sites are summarized in Table 10. It can be seen from this table that although groundwater level up to 2 lag months was considered as input variable for all the sites, only three sites (i.e., Sites E-5, F-6 and GH-4.5) show the influence of 2-month lag of groundwater level on current groundwater level. This is because in the stepwise regression method, the effect of 2-month lag was found to be insignificant based on *t*-value and *p*-level.

Comparative evaluation of the approaches used for MLR modeling

The MLR Models developed for individual sites using three approaches were compared in terms of statistical indicators such as *R*, Adjusted R^2, SEE, F-statistic and *p*-level as shown in Table 11. It is apparent from Table 11 that the MLR models developed using Approach III (i.e., considering three hydrological variables, 11 seasonal dummies and significant lags as inputs) have the highest values of multiple *R* and adjusted R^2 and the lowest value of SEE for all the 17 sites and also their significance is supported by corresponding F-statistic and *p*-level. Therefore, it can be inferred that the MLR models developed using Approach III provide the best results compared to the MLR models developed using Approach I and Approach II. Moreover, the comparison of the results of Approach II and Approach III (Table 11) indicates that inclusion of 2-month lags of rainfall, temperature, river stage and groundwater level in the regression model has a reasonable influence in the fluctuations of groundwater at 17 sites.

Moreover, a visual comparison of the observed and predicted groundwater levels using MLR models based on the three approaches was also performed. As example, Figs. 4, 5 show the simultaneous plots of observed and predicted groundwater levels at Sites B-3 and I-2 for the 1999–2004 period using three approaches. Similarly, the scatter plots of observed and predicted groundwater levels

Table 10 MLR models for predicting groundwater levels at 17 sites using Approach III

Site	MLR Models (Approach III)
A-2	$\mathrm{GWL} = 11.6764 + 0.0004 \times R + 0.0048 \times T + 0.7955 \times S + 0.2692 \times D_4 + 0.0242 \times D_5 + 0.0786 \times D_7 - 0.1190 \times D_9 - 0.0435 \times D_{10} + 0.2789 \times \mathrm{GW}_{t-1} + 0.0002 \times R_{t-1} - 0.4698 \times S_{t-1} + 0.0241 \times T_{t-1} - 0.0117 \times T_{t-2}$
B-3	$\mathrm{GWL} = 14.1116 + 0.0004 \times R + 0.0418 \times D_3 + 0.0380 \times D_4 + 0.0546 \times D_8 - 0.0360 \times D_9 + 0.2814 \times \mathrm{GW}_{t-1}\ 0.0002 \times R_{t-1} - 0.1610 \times S_{t-1} + 0.0777 \times S_{t-2} + 0.0039 \times T_{t-1} - 0.0038 \times T_{t-2}$
C-2	$\mathrm{GWL} = 6.1789 + 0.0014 \times R + 0.0683 \times T - 0.2486 \times D_1 + 0.3048 \times D_3 + 1.4176 \times D_4 - 0.6908 \times D_8 - 0.7069 \times D_9 + 0.4746 \times \mathrm{GW}_{t-1} - 0.0004 \times R_{t-2} - 0.4632 \times S_{t-1} + 0.0628 \times T_{t-1} - 0.0403 \times T_{t-2}$
C-7	$\mathrm{GWL} = 4.9791 - 0.0002 \times R + 0.0045 \times T + 0.0770 \times D_4 - 0.0686 \times D_8 - 0.0626 \times D_{11} + 0.3808 \times \mathrm{GW}_{t-1} - 0.0001 \times R_{t-1}$
D-6	$\mathrm{GWL} = 3.2901 + 0.0004 \times R + 0.0251 \times T + 0.5734 \times S - 0.2282 \times D_2 + 0.3631 \times D_4 - 0.2778 \times D_8 - 0.1481 \times D_9 + 0.3665 \times \mathrm{GW}_{t-1} + 0.0003 \times R_{t-1} - 0.5471 \times S_{t-1} + 0.0212 \times T_{t-1} - 0.0181 \times T_{t-2}$
E-2	$\mathrm{GWL} = 3.5354 + 0.0226 \times T + 1.1602 \times S - 0.4135 \times D_1 - 0.6927 \times D_2 + 0.9200 \times D_4 - 0.0394 \times D_5 + 0.5331 \times D_7 + 0.3278 \times D_{10} + 0.2083 \times D_{11} + 0.4135 \times \mathrm{GW}_{t-1} - 0.8281 \times S_{t-1} + 0.0471 \times T_{t-1} - 0.0487 \times T_{t-2}$
E-4	$\mathrm{GWL} = 3.3473 + 0.0013 \times R + 0.0323 \times T - 0.3304 \times D_1 - 0.6411 \times D_2 + 1.0090 \times D_4 + 0.4694 \times D_7 - 0.1602 \times D_8 - 0.2707 \times D_9 + 0.3455 \times \mathrm{GW}_{t-1} + 0.0004 \times R_{t-1} - 0.5543 \times S_{t-1} - 0.2283 \times S_{t-2} + 0.0744 \times T_{t-1} - 0.0419 \times T_{t-2}$
E-5	$\mathrm{GWL} = 2.8829 + 0.0009 \times R + 0.0098 \times T + 0.2904 \times S - 0.2769 \times D_2 + 1.0533 \times D_4 + 0.2419 \times D_6 + 0.9683 \times D_7 - 0.1157 \times D_9 - 0.1061 \times D_{11} + 0.4837 \times \mathrm{GW}_{t-1} + 0.0003 \times R_{t-1} - 0.1554 \times \mathrm{GW}_{t-2} - 0.6012 \times S_{t-1} + 0.0775 \times T_{t-1} - 0.0300 \times T_{t-2}$
F-1	$\mathrm{GWL} = 1.2098 + 0.0003 \times R - 0.0019 \times T + 0.9999 \times S + 0.1387 \times D_3 + 0.3847 \times D_4 + 0.0892 \times D_5 + 0.1942 \times D_7 + 0.1212 \times D_6 + 0.4679 \times \mathrm{GW}_{t-1} + 0.0002 \times R_{t-1} - 0.6921 \times S_{t-1} + 0.0203 \times T_{t-1} - 0.0073 \times T_{t-2}$
F-6	$\mathrm{GWL} = 1.9303 + 0.0005 \times R + 0.0068 \times T + 0.2200 \times S - 0.0624 \times D_1 - 0.2475 \times D_2 + 0.6562 \times D_4 + 0.1145 \times D_6 + 0.5142 \times D_7 + 0.5389 \times \mathrm{GW}_{t-1} - 0.1032 \times \mathrm{GW}_{t-2} - 0.0001 \times R_{t-2} - 0.2451 \times S_{t-1} + 0.0472 \times T_{t-1} - 0.0254 \times T_{t-2}$
G-2	$\mathrm{GWL} = 0.0353 + 0.0009 \times R + 0.0333 \times T + 0.3773 \times S + 0.3063 \times D_3 + 1.0182 \times D_4 - 0.4576 \times D_8 - 0.6415 \times D_9 - 0.3814 \times D_{10} - 0.1563 \times D_{11} + 0.2874 \times \mathrm{GW}_{t-1} + 0.0004 \times R_{t-1} - 0.6049 \times S_{t-1} + 0.0659 \times T_{t-1}$
GH-4.5	$\mathrm{GWL} = -1.0752 + 0.0010 \times R + 0.2571 \times D_1 + 0.4967 \times D_2 + 0.8587 \times D_3 + 0.9663 \times D_4 - 0.2646 \times D_9 + 0.3535 \times D_{11} + 0.6862 \times \mathrm{GW}_{t-1} - 0.1747 \times \mathrm{GW}_{t-2} - 0.5948 \times S_{t-1} - 0.2058 \times S_{t-2} + 0.0968 \times T_{t-1}$
H-2	$\mathrm{GWL} = -0.1482 + 0.0002 \times R + 0.0205 \times T + 0.2340 \times S + 0.0875 \times D_3 + 0.2630 \times D_4 - 0.0863 \times D_5 - 0.0878 \times D_8 - 0.1017 + 0.5715 \times \mathrm{GW}_{t-1} - 0.0003 \times R_{t-2} + 0.3510 \times S_{t-1} + 0.1970 \times S_{t-2} + 0.0077 \times T_{t-1}$
H-3	$\mathrm{GWL} = -0.4464 + 0.0008 \times R + 0.0245 \times T + 0.1090 \times D_2 + 0.2997 \times D_3 + 0.8362 \times D_4 - 0.3261 \times D_8 - 0.4537 \times D_9 - 0.2168 \times D_{10} + 0.4105 \times \mathrm{GW}_{t-1} + 0.0003 \times R_{t-1} + 0.4633 \times S_{t-1} + 0.0597 \times T_{t-1}$
H-4	$\mathrm{GWL} = -0.1953 + 0.0006 \times R + 0.0174 \times T + 0.2813 \times D_3 + 0.7017 \times D_4 + 0.1205 \times D_7 - 0.0770 \times D_9 + 0.3152 \times D_{11} + 0.5920 \times \mathrm{GW}_{t-1} + 0.0004 \times R_{t-1} - 0.0003 \times R_{t-2} - 0.5230 \times S_{t-1} - 0.0605 \times S_{t-2} + 0.0406 \times T_{t-1} - 0.0103 \times T_{t-2}$
H-5	$\mathrm{GWL} = -0.5839 + 0.0003 \times R + 0.1929 \times S + 0.0786 \times D_1 - 0.1764 \times D_2 + 0.3144 \times D_3 + 0.3337 \times D_4 + 0.0895 \times D_6 + 0.0742 \times D_{11} + 0.6097 \times \mathrm{GW}_{t-1} - 0.3771 \times S_{t-1} + 0.0360 \times T_{t-1}$
I-2	$\mathrm{GWL} = -0.0756 + 0.0132 \times T + 0.2486 \times S + 0.0848 \times D_4 - 0.0523 \times D_5 + 0.0486 \times D_{10} + 0.4213 \times \mathrm{GW}_{t-1} - 0.0001 \times R_{t-1} - 0.0002 \times R_{t-2} - 0.1347 \times S_{t-1} + 0.1191 \times S_{t-2} + 0.0031 \times T_{t-1}$

Table 11 Comparison of MLR models developed using three approaches at 17 sites

Site	MLR Model with hydrological variables (Approach I)					MLR Model with hydrological variables and seasonal dummies (Approach II)					MLR Model with hydrological variables, seasonal dummies and significant lags (Approach III)				
	R	Adj. R^2	SEE (m)	F-statistic	p-level	R	Adj. R^2	SEE (m)	F-statistic	p-level	R	Adj. R^2	SEE (m)	F-statistic	p-level
A-2	0.956	0.911	0.125	242.85	0.0000	0.972	0.939	0.103	183.14	0.0000	0.980	0.951	0.092	108.04	0.0000
B-3	0.829	0.678	0.061	75.70	0.0000	0.850	0.697	0.059	28.16	0.0000	0.897	0.768	0.052	22.39	0.0000
C-2	0.887	0.780	0.615	126.66	0.0000	0.952	0.894	0.428	67.22	0.0000	0.970	0.929	0.350	78.10	0.0000
C-7	0.400	0.135	0.082	6.56	0.0030	0.486	0.178	0.080	4.08	0.0030	0.660	0.373	0.070	7.04	0.0000
D-6	0.902	0.805	0.246	98.50	0.0000	0.943	0.875	0.196	63.21	0.0000	0.960	0.905	0.172	57.13	0.0000
E-2	0.898	0.800	0.445	143.23	0.0000	0.971	0.933	0.259	99.11	0.0000	0.981	0.954	0.214	113.67	0.0000
E-4	0.922	0.846	0.451	195.87	0.0000	0.973	0.938	0.286	120.35	0.0000	0.983	0.957	0.239	113.58	0.0000
E-5	0.941	0.882	0.375	266.03	0.0000	0.980	0.952	0.239	118.11	0.0000	0.989	0.971	0.184	162.25	0.0000
F-1	0.957	0.913	0.140	248.95	0.0000	0.970	0.934	0.122	144.67	0.0000	0.983	0.958	0.097	126.10	0.0000
F-6	0.946	0.892	0.233	293.66	0.0000	0.980	0.954	0.152	163.06	0.0000	0.990	0.975	0.112	198.49	0.0000
G-2	0.936	0.870	0.319	158.95	0.0000	0.979	0.949	0.199	110.58	0.0000	0.984	0.962	0.172	138.68	0.0000
GH-4.5	0.787	0.608	0.553	56.03	0.0000	0.935	0.854	0.338	42.43	0.0000	0.972	0.933	0.228	83.85	0.0000
H-2	0.892	0.789	0.158	134.00	0.0000	0.944	0.873	0.122	49.73	0.0000	0.968	0.923	0.095	66.83	0.0000
H-3	0.912	0.827	0.306	170.61	0.0000	0.980	0.953	0.160	144.28	0.0000	0.988	0.970	0.127	193.83	0.0000
H-4	0.861	0.734	0.310	99.03	0.0000	0.947	0.883	0.206	60.33	0.0000	0.980	0.951	0.133	99.59	0.0000
H-5	0.736	0.529	0.253	40.82	0.0000	0.927	0.834	0.150	33.33	0.0000	0.960	0.907	0.112	64.00	0.0000
I-2	0.892	0.790	0.094	134.11	0.0000	0.935	0.854	0.079	42.38	0.0000	0.957	0.900	0.065	58.86	0.0000

R Correlation coefficient, R^2 Coefficient of determination, *SEE* Standard error of estimate

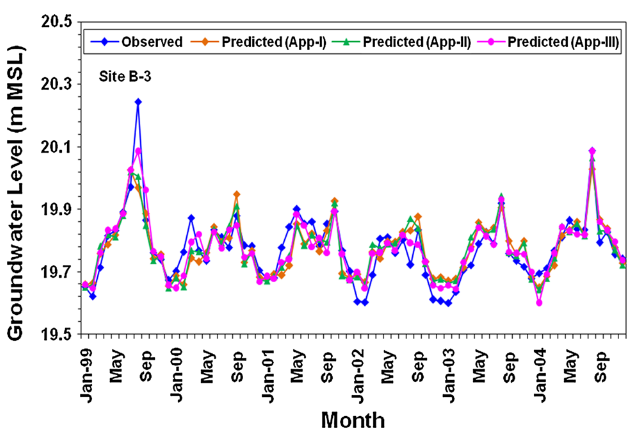

Fig. 4 Comparison of observed and predicted groundwater levels (Approaches I, II and III) at Site B-3 for the 1999–2004 periods

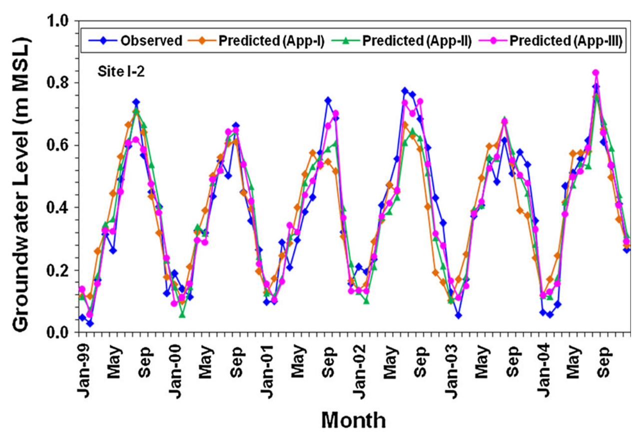

Fig. 5 Comparison of observed and predicted groundwater levels (Approaches I, II and III) at Site I-2 for the 1999–2004 periods

at Sites B-3 and I-2 following three approaches are illustrated in Figs. 6, 7. These plots also indicate that the prediction accuracy of the MLR models developed based on Approach III is mostly better than other two approaches.

Calibration and validation of best MLR models

The coefficients of MLR models developed at 17 sites considering three influential hydrological variables, 11

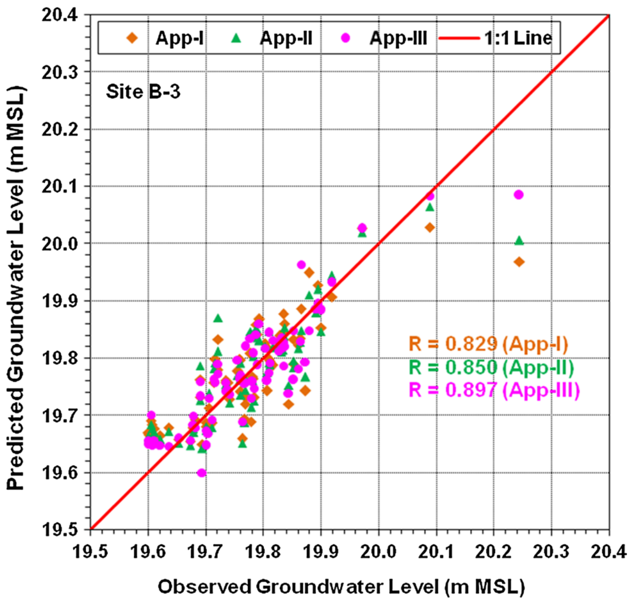

Fig. 6 Scatter plots of observed and predicted groundwater levels by MLR models using three approaches at Site B-3 for the 1999–2004 periods

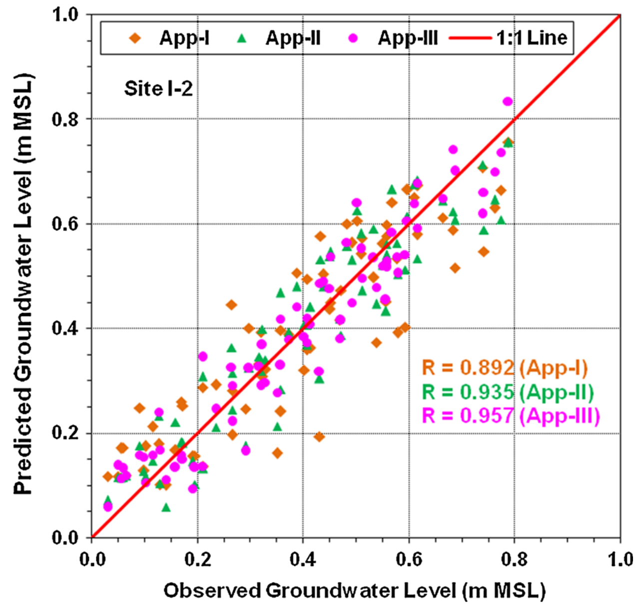

Fig. 7 Scatter plots of observed and predicted groundwater levels by MLR models using three approaches at Site I-2 for the 1999–2004 periods

seasonal dummy variables, and significant lags of hydrological variables and groundwater levels and using 1999–2002 data for calibration are presented in Table 12. The calibration and validation results for the MLR Models are presented in Table 13 in terms of statistical indicators such as R, multiple R^2 and SEE. It is apparent from Table 13 that the multiple R^2 values among the 17 sites vary from 0.780 (Site B-3) to 0.973 (Sites E-4 and G-2) for the calibration period and from 0.709 (Site D-6) to 0.938 (Site H-3) for the validation period which suggest that the developed MLR models can predict groundwater levels reasonably at all the sites. However, the MLR model of Site C-7 has the lowest multiple R^2 (0.544 for calibration and 0.007 for validation) among the 17 sites. This indicates that the groundwater level at Site C-7 is not influenced by the factors considered in the study and is affected by some anthropogenic factors such as the existence of a rubber dam very close to Site C-7 (Fig. 1). Moreover, the MLR models developed using the entire (6-year) datasets yield higher multiple R^2 and lower SEE compared to that using only calibration datasets for a majority of the sites, thereby indicating relatively better MLR models for the larger dataset. Further, a comparison of regression results between calibration and validation datasets showed lower values of multiple R^2 and higher values of SEE for validation datasets at all the sites as compared to that of calibration datasets.

Results of diagnostic checking of developed MLR models

After developing the MLR models, they were checked for the basic assumptions of multiple linear regression models for prediction. In Fig. 8, the plots of the standardized residuals (errors) and predicted values at Site-B-3 (using Approach-I) are shown to check the assumption of homoscedasticity. It is clear from this figure that the residuals are constant with respect to the horizontal (zero residual) line and there is no increasing spread of residuals from left to right, thereby satisfying the assumption of constant variance of errors.

The assumption of normality was checked by histogram of residuals and normal probability plots. Figure 9 shows the histogram of residuals with normal curve superimposed. In this figure, the residuals are distributed in such a manner that they resemble the normal distribution curve. Figure 10 shows the normal probability plots of residuals and it is clearly observed from this figure that majority of the residuals are falling on the straight line and hence it is proved that the residuals obtained are normally distributed.

Table 12 Regression coefficients of MLR models for predicting groundwater levels at 17 sites using calibration datasets

Site	Regression coefficients											
	Intercept	GW_{t-1}	GW_{t-2}	R	R_{t-1}	R_{t-2}	T	T_{t-1}	T_{t-2}	S	S_{t-1}	S_{t-2}
A-2	6.1572	–	–	0.0003	0.0003	–	0.0129	0.0133	–	0.8803	−0.2404	−0.0647
B-3	14.2696	0.2739	–	0.0004	0.0003	–	–	–	–	–	−0.1529	–
C-2	4.9457	0.5431	–	0.0011	–	−0.0007	0.1033	–	–	–	–	−0.4311
C-7	4.0716	0.4985	–	−0.0002	−0.0003	–	–	–	–	–	0.1618	0.0729
D-6	5.4199	–	–	0.0003	0.00358	–	0.0432	–	−0.0186	0.6289	–	–
E-2	2.9648	0.4037	–	–	0.0008	–	0.0235	0.0314	–	1.3327	−1.4295	–
E-4	2.6132	0.3776	0.0951	0.0011	0.0004	−0.0009	0.0748	–	−0.0248	–	–	–
E-5	2.0439	0.5285	–	0.0011	–	0.0005	0.0349	0.0421	−0.0421	–	–	–
F-1	2.3141	0.2649	−0.2179	0.0004	0.0003	0.0003	−0.0079	0.0357	0.0121	1.0911	−0.5139	–
F-6	1.6247	0.4951	–	0.0006	–	−0.0003	0.0252	0.0322	−0.0262	–	–	–
G-2	0.4252	0.2498	0.1022	0.0008	0.0005	−0.0006	0.0615	–	–	0.3063	0.2998	–
GH-4.5	−0.9839	0.647	−0.2079	0.001	0.0005	–	–	0.0902	–	–	−0.6591	–
H-2	−0.0774	0.5011	0.3718	0.0003	–	0.0004	0.0257	0.0097	−0.0199	0.0861	0.2208	–
H-3	−0.1759	0.3995	–	0.0007	0.0005	−0.0002	0.0261	0.0373	–	–	−0.3886	–
H-4	−0.0068	0.4849	−0.0927	0.0007	0.0005	–	–	0.0584	−0.0102	–	0.4554	–
H-5	−0.5338	0.5748	–	0.0003	–	–	–	0.0356	–	0.2817	−0.4258	–
I-2	−0.0626	0.4272	–	–	−0.0002	0.0004	0.0175	−0.0033	–	0.2311	–	0.1847

Site	Regression coefficients										
	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}
A-2	–	–	–	0.2878	0.0917	–	–	–	−0.104	−0.102	-
B-3	–	–	0.0375	–	–	–	–	0.0663	–	–	–
C-2	–	0.2836	0.6288	1.622	–	–	–	−0.7022	−0.3603	−0.4169	–
C-7	-	-	-	0.0519	-	0.0489	-	-	0.0666	–	–
D-6	−0.212	−0.5236	−0.3439	0.062	–	–	–	−0.163135	–	–	–
E-2	−0.2348	−0.2587	0.4599	1.4086	0.3248	–	0.4926	–	–	–	–
E-4	−0.3106	−0.5201	–	1.05	–	–	0.4598	−0.1631	–	–	0.1911
E-5	–	−0.2406	-	1.0076	-	-	0.9641	–	–	–	–
F-1	–	–	–	0.2895	0.0426	0.108	0.1486	–	–	–	–
F-6	–	−0.1702	–	0.6183	–	–	0.4527	–	–	−0.0948	–
G-2	0.182	−0.2577	–	0.8261	–	0.246	–	–	–	–	0.2432
GH-4.5	0.2445	0.4051	0.7339	0.9149	–	–	–	–	–	–	0.3509
H-2	–	–	–	0.115	−0.1496	−0.108	–	–	–	–	0.09
H-3	–	–	0.205	0.7763	–	0.0758	0.2191	–	–	–	0.21
H-4	–	–	0.2399	0.6511	–	–	0.1432	–	–	–	0.157
H-5	–	0.1116	0.2362	0.2652	−0.0859	–	−0.1053	–	–	–	–
I-2	–	–	–	–	0.0748	–	–	–	–	–	–

GW Groundwater level, *R* Rainfall, *T* Temperature, *S* River stage D_1, D_2, D_3,$\cdots D_{11}$ Seasonal dummy variables, *t−1* Lag of one month, *t−2* Lag of two months

Table 13 Goodness-of-fit statistics for the MLR models developed using calibration and validation datasets at 17 sites

Site	Calibration datasets			Validation datasets			Calibration + validation datasets		
	R	SEE (m)	Multiple R^2	R	SEE (m)	Multiple R^2	R	SEE (m)	Multiple R^2
A-2	0.981	0.090	0.963	0.956	0.179	0.914	0.980	0.092	0.960
B-3	0.884	0.057	0.780	0.858	0.061	0.736	0.897	0.052	0.805
C-2	0.973	0.351	0.947	0.926	0.671	0.858	0.970	0.350	0.941
C-7	0.738	0.067	0.544	0.083	0.104	0.007	0.660	0.070	0.436
D-6	0.969	0.163	0.939	0.842	0.346	0.709	0.960	0.172	0.922
E-2	0.944	0.327	0.891	0.964	0.388	0.930	0.981	0.214	0.962
E-4	0.986	0.227	0.973	0.942	0.576	0.887	0.983	0.239	0.966
E-5	0.975	0.207	0.950	0.963	0.352	0.928	0.989	0.184	0.978
F-1	0.957	0.189	0.917	0.942	0.241	0.888	0.983	0.097	0.966
F-6	0.940	0.207	0.885	0.888	0.213	0.789	0.990	0.112	0.980
G-2	0.986	0.174	0.973	0.954	0.391	0.910	0.984	0.172	0.968
GH-4.5	0.979	0.221	0.958	0.917	0.407	0.841	0.972	0.228	0.945
H-2	0.975	0.090	0.951	0.895	0.235	0.802	0.968	0.095	0.937
H-3	0.949	0.317	0.901	0.968	0.266	0.938	0.988	0.127	0.976
H-4	0.973	0.124	0.948	0.921	0.307	0.849	0.980	0.133	0.960
H-5	0.927	0.135	0.860	0.919	0.156	0.845	0.960	0.112	0.922
I-2	0.932	0.070	0.870	0.928	0.097	0.863	0.957	0.065	0.916

R Correlation coefficient, *SEE* Standard error of estimate, R^2 Coefficient of determination

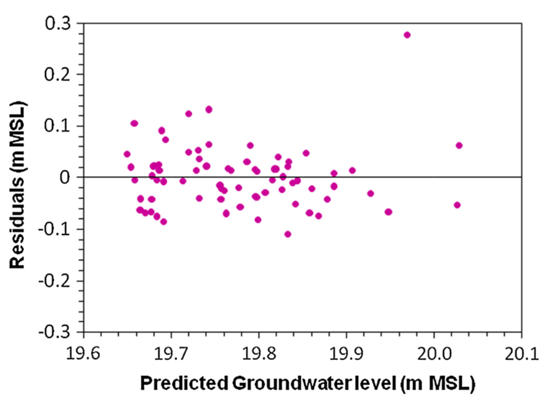

Fig. 8 Plots of residuals vs. predicted values for Site B-3 (Approach I)

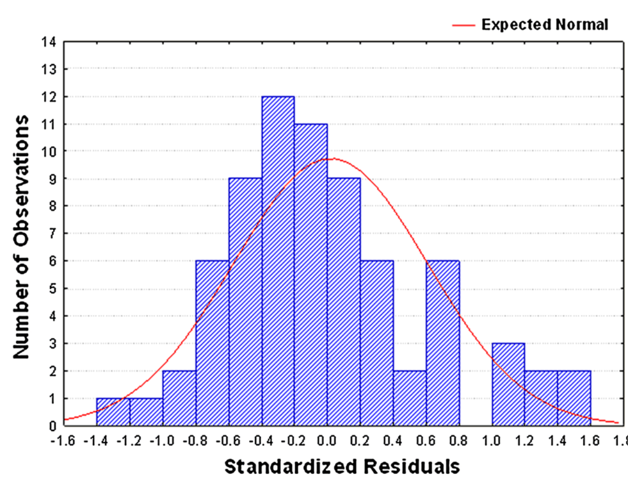

Fig. 9 Histogram of residuals for Site B-3 (Approach I)

The autocorrelation of residuals was checked by computing Durbin-Watson statistic. The value of Durbin-Watson statistic for the residuals obtained at Site B-3 is 1.58, which is significant at 95 % confidence level, thereby satisfying the condition of no autocorrelation at lag 1. Similarly, for all the 17 sites these diagnostic checks were employed for all approaches to test the basic assumptions of multiple linear regression technique.

Besides the basic assumptions, the degree of multicollinearity was also checked using variance inflation factor (V1F) and tolerance. As a rule of thumb, a variable with a VIF value of greater than 5 or tolerance <0.2 corresponds to high multicollinearity (Statistica 2001). As an example, Table 14 shows the check for multicollinearity at Site B-3 using Approach III which indicated that the magnitude of

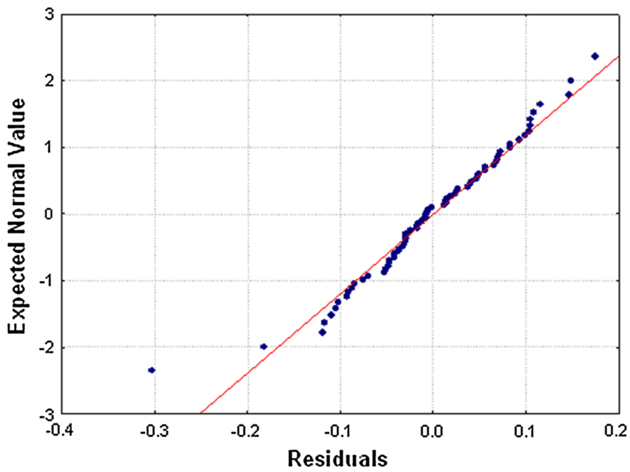

Fig. 10 Normal probability plot of residuals for Site B-3 (Approach I)

Table 14 Check for multicollinearity at Site B-3 using Approach III

Variables	R^2	Collinearity statistics		t (60)	p-Level
		Tolerance	VIF		
Rainfall (mm)	0.407	0.593	1.687	8.0942	0.0000
$GW_{(t-1)}$	0.711	0.289	3.464	2.6511	0.0102
D_9	0.335	0.665	1.504	−1.3317	0.0188
D_8	0.280	0.720	1.390	2.1030	0.0397
$Temp_{(t-2)}$	0.762	0.238	4.202	−1.7176	0.0091
$Temp_{(t-1)}$	0.770	0.230	4.347	1.6846	0.0073
D_3	0.301	0.699	1.431	1.5875	0.0012
D_4	0.292	0.708	1.413	1.4514	0.0052
River $stage_{(t-1)}$	0.746	0.254	3.935	−2.7365	0.0082
$Rainfall_{(t-1)}$	0.713	0.287	3.485	2.3711	0.0121
River stage $_{(t-2)}$	0.670	0.330	3.027	1.9395	0.0254

R^2 Coefficient of determination, *VIF* Variance inflation factor, *GW* Groundwater level, D_3, D_4, D_8, D_9 Seasonal dummy variables, $t-1$ Lag of one month, t *2* Lag of two months

VIF is less than 5 and also tolerance is greater than 0.2 for all independent variables included in the regression model, and hence no multicollinearity. Similarly, for all the 17 sites the collinearity diagnostic check was employed and found the absence of multicollinearity in the independent variables.

Conclusions

This study demonstrates a robust methodology to develop efficient MLR models for forecasting groundwater levels. The MLR models were developed considering influential environmental factors having significant effects on groundwater levels. These inputs/factors were selected by three approaches based upon plausible combinations of significant inputs. For the identification of significant inputs for the MLR models, the stepwise regression technique and partial autocorrelation technique were found superior to the standard regression technique and autocorrelation technique, respectively. Although better correlation between monthly temperature and monthly groundwater levels is attributed to the less temporal variability of these two parameters in the study area, further investigation in this direction is necessary. The MLR models developed using three approaches were evaluated using statistical as well as graphical methods. Of the three approaches, the MLR models developed using Approach III considering all possible but significant inputs (rainfall, temperature, river stage, seasonal dummy variables and significant lags) were found to be the best for predicting monthly groundwater levels at individual sites, followed by the MLR models based on Approach II. Further, the analysis of MLR models based on significant lags and seasonal dummy variables revealed that the seasonal dummy variables as well as the lags of rainfall, temperature and river stage up to 2 months have a significant influence on the groundwater levels of all the sites. The calibration and validation results of the best MLR models revealed that they are capable of predicting groundwater levels with a reasonable accuracy (Multiple R^2 ranging 0.780–0.973 during calibration and from 0.709 to 0.938 during validation) at almost all the sites under the study. The significant advantage of MLR models is that they can provide dependable prediction from a limited record of field data.

The findings of this study are of great importance for improving planning and management of groundwater resources, where developing numerical groundwater models is a serious issue due to data uncertainty and data deficiency. The use of MLR models as a prediction tool is much easier as they require minimal information for development, less time-consuming and the results can be interpreted in a simpler manner, which make MLR models more attractive and cost-effective for modeling groundwater levels. The methodology presented in this study can be easily replicated in other hydrogeologic settings of the world, especially in the developing nations where data scarcity and quality pose a great hindrance to the application of process-based models.

References

Abdelgawad AM, Watanabe K, Mohammed M, Takeuchi S (2010) Estimating groundwater residence time using multiple regression models based on fluoride dissolution: an exploration of possibilities. Environ Earth Sci 60(3):449–462

Adeloye AJ (2009) Multiple linear regression and artificial neural network models for generalized reservoir storage–yield–

reliability function for reservoir planning. J Hydrologic Eng, ASCE 14(7):731–738

Ahsan SMF, Islam MN, Uddin MJ, Uddin MT (2008) Statistical modeling of groundwater arsenic contamination level in Bangladesh due to chemical elements. J Appl Quant Methods 3(3):254–262

Bayramov E, Buchroithner MF, McGurty E (2012) Determination of main climate and ground factors controlling vegetation cover regrowth along oil and gas pipelines using multiple, spatial and geographically weighted regression procedures. Environ Earth Sci 66(7):2047–2062

Boy-Roura M, Nolan BT, Menció A, Mas-Pla J (2013) Regression model for aquifer vulnerability assessment of nitrate pollution in the Osona region (NE Spain). J Hydrol 505:150–162

Chen H, Druliner D (1988) Agricultural chemical contamination of ground water in six areas of the high plains aquifer, Nebraska, National Water Summary 1986-Hydrologic events and ground water quality. Water-Supply Paper 2325, Reston, Virginia, US. Geological Survey

Coppola EA, Szidarovszky F, Poulton MM, Charles E (2003) Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions. J Hydrologic Eng, ASCE 8(6):348–360

Coppola EA, Rana AJ, Poulton MM, Szidarovszky F, Uhl VW (2005) A neural network model for predicting aquifer water level elevations. Ground Water 43(2):231–241

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

Cui Y, Du X (2011) Soil heavy-metal speciation and wheat phytotoxicity in the vicinity of an abandoned lead–zinc mine in Shangyu City, eastern China. Environ Earth Sci 62(2):257–264

Daliakopoulos IN, Coulibaly P, Tsanis IK (2005) Groundwater level forecasting using artificial neural network. J Hydrol 309:229–240

Eldaw AK, Salas JD, Garcia LA (2003) Long-range forecasting of the Nile river flows using climatic forcing. J Appl Meteorol 42(7):890–904

Heuvelmans G, Muys B, Feyen J (2006) Regionalisation of the parameters of a hydrological model: comparison of linear regression models with artificial neural nets. J Hydrol 319:245–265

Hodgson FDI (1978) The use of multiple linear regression in simulating ground-water level responses. Ground Water 16(4):249–253

Hsieh WW, Yuval LJ, Shabbar A, Smith S (2003) Seasonal prediction with error estimation of Columbia River streamflow in British Columbia. J Water Resour Plan Manag, ASCE 129(2):146–149

Jha MK, Sahoo S (2014) Efficacy of neural network and genetic algorithm techniques in simulating spatio-temporal fluctuations of groundwater. Hydrological Processes. doi:10.1002/hyp.10166

Jha MK, Chikamori K, Kamii Y, Yamasaki Y (1999) Field investigations for sustainable groundwater utilization in the Konan basin. Water Resour Manag 13:443–470

Jha MK, Jayalekshmi K, Machiwal D, Kamii Y, Chikamori K (2004) Determination of hydraulic parameters of an unconfined alluvial aquifer by the floodwave-response technique. Hydrogeol J 12(6):628–642

Kayabasi A (2012) Prediction of pressure meter modulus and limit pressure of clayey soils by simple and non-linear multiple regression techniques: a case study from Mersin, Turkey. Environ Earth Sci 66(8):2171–2183

Liu B, Zhao W, Zeng F (2013) Statistical analysis of the temporal stability of soil moisture in three desert regions of northwestern China. Environ Earth Sci 70(5):2249–2262

Makridakis S, Wheelwright S, Hyndman RJ (2008) Forecasting methods and applications, 3rd edn. John Wiley and Sons (Asia) Ltd, Singapore, p 656

McCuen RH, Rawls WJ, Whaley BL (1979) Comparative evaluation of statistical methods for water supply forecasting. Water Resour Bull 15(4):935–947

Nolan BT, Hitt KJ (2006) Vulnerability of shallow groundwater and drinking water wells to nitrate in the United States. Environ Sci Technol 40(24):7834–7840

Nourani V, Mogaddam AA, Nadiri AO (2008) An ANN-based model for spatiotemporal groundwater level forecasting. Hydrol Process 22:5054–5066

Risley JC, Gannett MW, Lea JK, Roehl EA (2005) An analysis of statistical methods for seasonal flow forecasting in the Upper Klamath River Basin of Oregon and California. Scientific Investigations Report No. 05-5177, USGS, Reston, VA 20192

Sahoo S, Jha MK (2013) Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment. Hydrogeol J. doi:10.1007/s10040-013-1029-5

Shao QX, Campbell NA (2002) Modelling trends in groundwater levels by segmented regression with constraints. Aust NZ J Stat 44(2):129–141

Siddiqui FI, Osman SBABS (2013) Simple and multiple regression models for relationship between electrical resistivity and various soil properties for soil characterization. Environ Earth Sci 70(1):259–267

Sinnakaudan SK, Ghani AA, Ahmad MSS, Zakaria NA (2006) Multiple linear regression model for total bed material load prediction. J Hydraulic Eng, ASCE 132(5):521–528

Spiegel MR, Stephens LJ (2008) Schaum's Outlines Statistics, 4th edn. Tata McGraw-Hill Publishing Company Ltd., New Delhi, p 577

Stackelberg PE, Gilliom RJ, Wolock DM, Hitt KJ (2006) Development and application of a regression equation for estimating the occurrence of atrazine in shallow ground water beneath agricultural areas of the United States: U.S. Geological Survey Scientific Investigations Report 2005-5287, p 12, online only

Stackelberg PE, Barbash JE, Gilliom RJ, Stone WW, Wolock DM (2012) Regression models for estimating concentrations of atrazine plus deethylatrazine in shallow groundwater in agricultural areas of the United States. J Environ Qual 41(2):479–494

Statistica (2001) STATISTICA: system reference. StatSoft, Inc., Tulsa, p 1098

Steichen J, Koelliker J, Grosh D, Heiman A, Yearout R, Robbins V (1988) Contamination of farmstead wells by pesticides, volatile organics, and inorganic chemicals in Kansas. Ground Water Monit Rev 8:153–160

Stigter TY, Ribeiro L, Carvalho Dill AMM (2008) Factorial regression models of groundwater contamination by agricultural practices: a case study in the south of Portugal. J Hydrol 357:42–56

Todd DK (1980) Groundwater Hydrology, 2nd edn. Wiley, New York, p 535

Tootle GA, Singh AK, Piechota TC, Farnham I (2007) Long lead-time forecasting of U.S. streamflow using partial least squares regression. J Hydrologic Eng, ASCE 12(5):442–451

Trichakis IC, Nikolos IK, Karatzas GP (2009) Optimal selection of artificial neural network parameters for the prediction of a karstic aquifer's response. Hydrol Process 23(20):2956–2969

Uddameri V (2007) Using statistical and artificial neural network models to forecast potentiometric levels at a deep well in South Texas. Environ Geol 51:885–895

USDA-NRCS (2007) Statistical techniques used in the VIPER water supply forecasting software. NRCS Technical Note No. 210-2, Engineering-Snow Survey and Water Supply Forecasting, http://directives.sc.egov.usda.gov/tn_sswsf_2_a.pdf (accessed on 25 August 2010)

Xu J, Peng S, Wang W, Yang S, Wei Q, Luo Y (2012) Prediction of daily reference evapotranspiration by a multiple regression method based on weather forecast data. Archi Agron Soil Sci, 1–15. doi:10.1080/03650340.2012.727400