Multivariate analysis of the heterogeneous geochemical processes controlling arsenic enrichment in a shallow groundwater system

Shuangbing Huang a, Changrong Liu a, Yanxin Wang b & Hongbin Zhan c

a Institute of Hydrogeology and Environmental Geology, Chinese Academy of Geological Sciences, Shijiazhuang, PR China
b State Key Laboratory of Biogeology and Environmental Geology and School of Environmental Studies, China University of Geosciences, Hubei, Wuhan, PR China
c Department of Geology and Geophysics, Texas A&M University, College Station, Texas, USA

Published online: 17 Dec 2013.


To link to this article: http://dx.doi.org/10.1080/10934529.2014.854689
Multivariate analysis of the heterogeneous geochemical processes controlling arsenic enrichment in a shallow groundwater system

SHUANGBING HUANG1, CHANGRONG LIU1, YANXIN WANG2 and HONGBIN ZHAN3

1Institute of Hydrogeology and Environmental Geology, Chinese Academy of Geological Sciences, Shijiazhuang, PR China
2State Key Laboratory of Biogeology and Environmental Geology and School of Environmental Studies, China University of Geosciences, Hubei, Wuhan, PR China
3Department of Geology and Geophysics, Texas A&M University, College Station, Texas, USA

The effects of various geochemical processes on arsenic enrichment in a high-arsenic aquifer at Jianghan Plain in Central China were investigated using multivariate models developed from combined adaptive neuro-fuzzy inference system (ANFIS) and multiple linear regression (MLR). The results indicated that the optimum variable group for the ANFIS model consisted of bicarbonate, ammonium, phosphorus, iron, manganese, fluorescence index, pH, and siderite saturation. These data suggest that reductive dissolution of iron/manganese oxides, phosphate-competitive adsorption, pH-dependent desorption, and siderite precipitation could integrally affect arsenic concentration. Analysis of the MLR models indicated that reductive dissolution of iron(III) was primarily responsible for arsenic mobilization in groundwaters with low arsenic concentration. By contrast, for groundwaters with high arsenic concentration (i.e., > 170 µg/L), reductive dissolution of iron oxides approached a dynamic equilibrium. The desorption effects from phosphate-competitive adsorption and the increase in pH exhibited arsenic enrichment superior to that caused by iron(III) reductive dissolution as the groundwater chemistry evolved. The inhibition effect of siderite precipitation on arsenic mobilization was expected to exist in groundwater that was highly saturated with siderite. The results suggest an evolutionary dominance of specific geochemical process over other factors controlling arsenic concentration, which presented a heterogeneous distribution in aquifers.

Supplemental materials are available for this article. Go to the publisher’s online edition of the Journal of Environmental Science and Health, Part A, to view the supplemental file.

Keywords: Arsenic enrichment, geochemical processes, heterogeneous, statistical models.

Introduction

Reductive dissolution of iron (III) oxides, hydroxides, and oxyhydroxides, collectively called in this study as iron oxides, has been proposed to be the dominant mechanism for arsenic mobilization.1-4 However, a weak positive correlation between arsenic and iron and/or other relevant redox-sensitive elements is commonly observed among numerous arsenic-contaminated groundwater studies.5-8 A recent study on the Jianghan Plain (JHP) in Central China has also reported a weak correlation between arsenic and redox-sensitive elements, although arsenic mobilization has been linked to the microbial reduction of iron oxides with labile dissolved organic matter (DOM) as electron donors and humic quinones as electron shuttles.9 Moreover, several groundwater samples contain abnormally high arsenic concentrations. These groundwater do not exhibit the general distribution found in most samples in terms of the relationship between arsenic and iron (Fig. S1; see online supplementary materials).

High arsenic groundwater is commonly heterogeneously distributed in aquifers.3, 10 Thus, arsenic concentration could be jointly controlled by multiple geochemical processes with varied degrees of influence for large-scale groundwater geochemistry investigation. In addition to the reductive dissolution of iron oxides, other geochemical processes could also simultaneously pose constraints on arsenic enrichment, which may reduce the influence of a single geochemical factor. However, this joint effect has not been explicitly discussed by previous studies despite its
Geochemical processes controlling arsenic enrichment in groundwater

This heterogeneous multi-factor system is difficult to illustrate accurately because the large-scale investigation of this system involves complex and nonlinear relations between geochemical factors and arsenic concentration. Although this topic needs further in-depth research, the existing knowledge on arsenic geochemistry are sufficient to aid in the identification of potential geochemical factors responsible for arsenic enrichment and facilitate the development of robust statistical models. Statistically, if multiple geochemical statistic variables can be properly fitted in a model, then the selected geochemical factors should correlate with the target variables (in this case, arsenic concentration). Thus, these factors could be considered as coexisting factors associated with arsenic occurrence and could be used to analyze further the potential geochemical process in multivariate models.

Multiple linear regression (MLR), one of the most popular regression techniques, employs a function that can better explain multivariate relationships compared with other techniques. MLR has been utilized to develop statistical model for complex As/F-contaminated groundwater systems.[14,15] However, MLR has limitations when solving nonlinear problems.[16] By contrast, the adaptive neuro-fuzzy inference system (ANFIS), which is a hybrid model of neural networks with fuzzy logics, is a powerful tool for nonlinear system identification because of its versatility in treating uncertainties and its ability to learn and adapt to a set of input and output data.[17,18]

However, ANFIS still cannot fully explain the internal relationship between independent and dependent variables because of its intrinsic implicit algorithm.[16] In this study, MLR and ANFIS were utilized to facilitate the interpretation of the effects of various geochemical processes on arsenic enrichment at JHP. ANFIS was employed to identify the correlation variables, and MLR was used to gather meaningful information to determine the relationships between arsenic concentrations and independent variables. To the best of our knowledge, very few studies have integrated the strengths of these two methods to analyze the joint geochemical effects on the arsenic contamination of groundwater systems.

JHP is a typical Quaternary alluvial-lacustrine river basin located at the middle reaches of the Yangtze River in Central China. The study area is located in the central part of the plain (Fig. 1), which is surrounded by the Han River, the Dongjing River (the tributary of the Han River), and the Yangtze River (Fig. 1). This low and flat central plain, which is crisscrossed by rivers and lakes, is one of the major agricultural and fishing areas in Central China. The surface sediments in this region include alluvial, alluvial-lacustrine, lacustrine-swamp, and lacustrine deposits (Fig. 1). Lacustrine deposits are characterized by abundant organic matter. High arsenic groundwater is typically found in a moderate reducing environment and contains high concentrations of Fe$^{2+}$, bicarbonate (HCO$_3^-$), and dissolved organic carbon. Moreover, the variations in DOM composition indicate the microbial mediation in the reductive dissolution of iron oxides with the oxidation of organic carbon.[9]

In this study, a hydrochemical data set from a field investigation of arsenic-contaminated groundwater was used to build statistical models using MLR and ANFIS. We validated mathematically the existence of multiple geochemical processes on arsenic enrichment. We also interpreted the various potential geochemical processes that control arsenic enrichment in groundwater.

Materials and methods

Data set

The data set for this study was obtained from the hydrogeochemical investigation of high arsenic groundwater at JHP by our group from July 26, 2011, to August 10, 2011. A total of 97 sets of water chemical data (see supplementary materials) were collected for the current study. The selected chemical parameters included major elements and redox-sensitive elements. Chemical analyses and quality control were performed according to a previously described procedure.[9] Three-dimensional fluorescence spectra of the DOMs were measured to determine the specific DOM functionality on the arsenic release process. Ninety-seven fluorescent excitation–emission matrices were collected through the method presented by Li et al.[19] Fluorescence index (FI) was designated as the indicator to distinguish DOM components from terrestrial or microbial sources.[20,21] This index was determined as the ratio of the emission intensity at a wavelength of 470 nm to that of the 520 nm wavelength obtained with an excitation of 370 nm.[21]

\[
FI = \frac{F_{470}}{F_{520}}
\]

(1)

where $F_{470}$ and $F_{520}$ are the emission intensities at 470 and 520 nm, respectively, both with excitations at 370 nm.

The saturation index (SI) with respect to a specific mineral was also added into the available database using the following formula:

\[
SI = \log(IAP/K_{sp}) = \log IAP - \log K_{sp}
\]

(2)

where $IAP$ represents the ion activity product, and $K_{sp}$ is the solubility of the product for a given temperature. $IAP = K_{sp}$ (SI = 0) indicates that the solution is at thermodynamic equilibrium with respect to a specific mineral. When $IAP > K_{sp}$, the solution is supersaturated and vice versa.
**Mathematical and assessment methods**

**ANFIS.** With a given input–output data set, a fuzzy inference system (FIS) with specific membership functions and “if–then” rules can be built using ANFIS. The membership functions can be determined using a corresponding set of parameters, which can then be tuned according to the learning procedure from the data being modeled, i.e., the backward propagation (BP) of neural networks known as “BP” algorithm.[18] In this study, we developed an ANFIS using a Sugeno-type FIS.[17] The output variable, which is the arsenic concentration in this case, of a Sugeno-type FIS is considered a parameterized nonlinear combination of the input variables (geochemical indices) based on predefined rules. A matrix \( Y (N \times D) \) is assumed, where \( N \) is the number of samples (referred to as data points), and \( D \) is the number of independent variables. A Sugeno-type fuzzy model with \( M \) rules (\( R_j, j = 1, 2, \ldots, M \)) can then be expressed in the following form:

\[
R_j: \text{If } y_i \in A_{ij} \text{ Then } z_j = f_j w_j
\]

where \( A_{ij} \) represents the selected fuzzy sets for the input variables \( y_i \) in the \( j \)-th rule, \( w_j \) is the firing strength, i.e., a weight indicating the effect of each rule on the prediction, and \( f_j \) is a linear function, which is a consequent function for the \( j \)-th rule. The fuzzy sets \( A_{ij} \) for the different rules are defined by subtractive clustering algorithm.[18] The resulting function is defined for each data point and is given as follows:

\[
f_j = a_0 + \sum_{i=1}^{D} a_i y_i \quad j = 1, \ldots, M
\]

where \( a_i \) represents the consequent parameters that must be optimized. The firing strength for each rule can be calculated by the “min” or “product” operator of the membership grades for each rule as follows:

\[
w_j = \min \left[ \mu_{ji} (y_i) \right] \quad (5)
\]

\[
w_j = \prod_{i=1}^{D} \mu_{ji} (y_i) \quad (6)
\]

where \( \mu_{ji} \) is the membership grade resulting from the membership function that describes the \( i \)-th variable in the \( j \)-th rule. The membership grade depends on the type and parameter of the membership function. Thus, we chose the most frequently used Gaussian membership function, which is expressed as follows:

\[
\mu_{ij} = \exp \left( -\frac{(y_i - c_{ij})^2}{2\sigma_{ij}^2} \right) \text{ for } i = 1, \ldots, D, j = 1, \ldots, M
\]

where parameters \( C_{ij} \) and \( \sigma_{ij} \) are the nonlinear premise parameters of the membership functions to be optimized.

The final output for each data point is then calculated as a normalized weighted average of the outcome of the different rules.

\[
\bar{Z} = \sum_{j=1}^{M} f_j \bar{w}_j \quad (8)
\]

where \( \bar{w}_j \) is the normalized activation function expressed as follows:

\[
\bar{w}_j = \frac{w_j}{\sum_{j=1}^{M} w_j} \quad (9)
\]

Two steps are undertaken when developing an ANFIS. First, the fuzzy sets that describe the different rules are...
Geochemical processes controlling arsenic enrichment in groundwater

481

constructed. Second, the premise and consequent parameters are optimized. A combination of gradient descent and least square method was applied to optimize the linear consequent and nonlinear premise parameters of the FIS. In this method, the nonlinear parameters (premise) were initially set at fixed values, and the consequent parameters were estimated using the least squares method. Then, the premise parameters were updated by a gradient descent optimization algorithm. This procedure was repeated until a specific optimization criterion was reached.\[17,18\]

**MLR.** Assuming that a linear relationship exists between the dependent variable and each predictor, then this relationship can be described as follows:

\[ y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + e_i \]  

(10)

where \( y_i \) is the value of the \( i \)-th case of the dependent scale variable, \( p \) is the number of predictors, \( b_j \) is the value of the \( j \)-th coefficient, \( j = 0, \ldots, p \), \( x_{ij} \) is the value of the \( i \)-th case of the \( j \)-th predictor, and \( e_i \) is the error in the observed value for the \( i \)-th case.

The model is considered linear because an increase in the value of the \( j \)-th predictor by one unit increases the value of the dependent by \( b_j \) units. \( b_0 \) is the intercept, which is the model-predicted value of the dependent variable when the value of every predictor is equal to zero.

**Statistical assessment.** Statistical indices, including the coefficient of determination \( (R^2) \), root mean square error (RMSE), and standard error of the estimate, were used to assess the model performance. The efficiency/response of the model output for ANFIS was measured using \( R^2 \) and RMSE.

1. \( R^2 \) ranges between 0 and 1. Higher values indicate better performance of the model.

\[ R^2 = 1 - \frac{\sum (y_{io} - y_{ic})^2}{\sum y_{io}^2 - \sum y_{ic}^2} \]  

(11)

2. RMSE indicates the discrepancy between the observed and calculated values. A lower RMSE corresponds to a more accurate prediction.

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{io} - y_{ic})^2} \]  

(12)

where \( y_{io} \) and \( y_{ic} \) denote the observed and calculated values, respectively, and \( n \) is the number of observations.

For the MLR model, similar to that of ANFIS, \( R^2 \) was also employed to reflect the accuracy of the entire regression goodness of fit, and the correlation coefficient \( R \) was simultaneously presented. Statistical measures, such as the standard error of the estimate, which is an index for evaluating the fitting accuracy of the model, and the significance level parameter \( P \) of the T test, which is used to assess the regression variables, were also used to evaluate further the model performance and the significance level of specific independent variables.

3. The standard error of the estimate can be expressed as follows:

\[ S_{y,x} = \sqrt{\frac{\sum (y_{io} - y_{ic})^2}{n - m - 1}} \]  

(13)

where \( S_{y,x} \) denotes the standard error of the estimate, \( y_{io} \) and \( y_{ic} \) are the observed and calculated values, respectively, \( n \) is the number of observations, and \( m \) is the number of variables. Thus, \( n-m-1 \) represents the degree of freedom of the standard error of the estimate.

**Model development procedure**

We preliminary identified the potential geochemical factors responsible for the high arsenic concentrations in the groundwater based on the knowledge-based understanding of arsenic geochemistry. Then, the identified geochemical parameters were selected as potential influencing variables (PIVs) of arsenic concentration and were applied into the ANFIS models to obtain the correlating variables with arsenic. Finally, the validated optimum variable group for the ANFIS model was used to develop various MLR models to geochemically interpret the arsenic enrichment. The model development process is presented in Fig. 2.

**Knowledge-based identification of geochemical factors.** The reduced species, such as Fe\(^{2+}\), dissolved sulfur, and NH\(_4\)\(^+\), are common accompanying products in reducing arsenic-rich groundwater.\[1,3\] A statistical histogram of redox-sensitive parameters shows an extensive occurrence of these products (Fig. 3). The ubiquity of NH\(_4\)\(^+\) can indicate the reducing levels of the aquifers qualified for Fe(III) reduction. Furthermore, the concentration of dissolved iron (reported as Fe\(^{2+}\)) with skewed distribution (Fig. 3) has a median of up to 7 mg/L, which demonstrates that Fe(III) is a major electron acceptor during redox reactions in aquifers. We also considered other sensitive chemical species, such as Mn\(^{2+}\), HCO\(_3\)\(^-\), and NH\(_4\)\(^+\), as variables associated with arsenic enrichment based on their hydrochemical characteristics. We added an FI to determine the effects of DOMs from specific sources on arsenic mobilization. We disregarded dissolved sulfur as a PIV because of its very low concentration in most of the samples (Fig. 3).

A number of studies have reported on the coprecipitation effect of pyrite and siderite with arsenic.\[22-24\] We considered siderite as a potential species that probably influences aqueous arsenic concentration through coprecipitation because HCO\(_3\)\(^-\) acts as the dominant anion with a high concentration of approximately 600 mg/L. The effect of pyrite was not considered because the presence of sulfur was substantially limited by the high concentration of dissolved iron in the groundwater. Given the competitive adsorption effect, we considered phosphate, which is
Fig. 2. Model development procedures and modeling objectives.

reported as total dissolved phosphorus (P) in the current study, as a PIV. In addition, pH was also considered a PIV for arsenic enrichment because of its key involvement in multiple hydrogeochemical processes.

Thus, we considered a total of eight preferable chemical parameters, namely, \( \text{HCO}_3^-, \text{NH}_4^+, \text{P}, \text{Fe}^{2+}, \text{Mn}^{2+}, \text{FI}, \text{pH} \), and the saturation of siderite, as PIVs closely associated with arsenic enrichment.

**ANFIS model.** ANFIS has a robust training and prediction capacity. Thus, each PIV correlated with arsenic concentration should theoretically achieve a good modeling result. The modeling objective of ANFIS was to examine whether the calculated arsenic concentration matched well with the observed value. Matching results demonstrated that arsenic concentration can be expressed by the ANFIS model with selected geochemical parameters, thereby validating the various geochemical control effects represented by these parameters.

To test our hypothesis, we used additional five sets of geochemical parameters as parallel model inputs by modifying the preliminary eight PIVs. Six sets of input variables were assigned to the ANFIS model, and the performances of the corresponding models were evaluated. The models are referred to as preferably selected eight-variable (PS8), preferably selected five-variable (PS5), moderately hybridized eight-variable (MH8), moderately hybridized five-variable (MH5), hyper-hybridized eight-variable (HH8), and hyper-hybridized five-variable (HH5) models. Detailed
Geochemical processes controlling arsenic enrichment in groundwater

Table 1. Specification on the types of variable group corresponding to model input variables.

<table>
<thead>
<tr>
<th>Types of variable combination</th>
<th>Corresponding geochemical parameters selected for model inputs</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS8</td>
<td>HCO$_3^-$, NH$<em>4^+$, P, Fe$^{2+}$, Mn$^{2+}$, FI, pH, SI$</em>{siderite}$</td>
<td>All symbols in second column represent corresponding dissolved species or probable ions in groundwater except FI, pH, SI$_{siderite}$.</td>
</tr>
<tr>
<td>PS5</td>
<td>HCO$<em>3^-$, P, Fe$^{2+}$, pH, SI$</em>{siderite}$</td>
<td></td>
</tr>
<tr>
<td>MH8</td>
<td>HCO$_3^-$, Fe$^{2+}$, FI, pH, SO$_4^-$, NO$_3^-$, Ca$^{2+}$, Na$^+$</td>
<td></td>
</tr>
<tr>
<td>MH5</td>
<td>HCO$_3^-$, Fe$^{2+}$, pH, Ca$^{2+}$, Na$^+$</td>
<td></td>
</tr>
<tr>
<td>HH8</td>
<td>Cl$^-$, NO$_3^-$, SO$_4^{2-}$, HCO$_3^-$, Al$^{3+}$, Ba$^{2+}$, Ca$^{2+}$, Na$^+$</td>
<td></td>
</tr>
<tr>
<td>HH5</td>
<td>HCO$_3^-$, Al$^{3+}$, Ba$^{2+}$, Ca$^{2+}$, Na$^+$</td>
<td></td>
</tr>
</tbody>
</table>

descriptions regarding these variable groups are presented in Table 1.

MLR model. MLR was used to gather information for the geochemical interpretation of arsenic enrichment through statistical assessment of a series of models. The intrinsic non-linear nature may reduce the suitability of the linear model during prediction, because the samples were obtained in a large scale to understand the regional geochemical processes of the system. Thus, a relatively significant estimate error may inevitably be introduced in the resulting model. However, this study aimed to derive a good interpretation of geochemical effects based on the correlation level of multiple PIVs with arsenic rather than on that of single variable with arsenic with common binary variable analysis. Therefore, the model prediction precision is not a major consideration.

The validated optimum variable group for the ANFIS model was first introduced to the MLR model to obtain the initial MLR model. Statistically, samples with significant deviations from the general distribution of most samples in a variable relationship (Fig. S1; see online supplementary materials) were considered as outliers and removed. MLR employs a mathematical screening method called casewise diagnostics, which chooses cases as outliers if their estimated error exceeds $n$ times (in this case, $n = 2$) of the standard deviation of the observed values. Therefore, models with all samples as cases and without outliers were run to create a contrast against the regression coefficients. Then, a forward selection stepwise regression technique$^{[14,15]}$ was employed to determine the statistically significant independent variables for the MLR models. Thus, the obtained model is the statistical optimum regression model. A number of added regression models with reselected samples as cases were used to analyze local geochemical phenomena based on specific objectives (Fig. 2).

Results

ANFIS model

ANFIS models with various variable combinations as inputs were obtained, except for HH5 because of a failure in model convergence. The correlation between the calculated and observed arsenic concentrations, as well as the $R^2$ values of the different models, are presented in Fig. 4. The correlation exhibited a consistent modeling performance in the following order: PS8, PS5, MH8, MH5, HH8, and HH5. This trend was also observed from the comparisons of the calculated values with the observed values (Fig. 5) and the statistical parameters (Table 2).

The regression goodness of fit of the variable groups could be ranked from highest to lowest, as follows: PS8, PS5, MH8, MH5, and HH8. This ranking corresponds well with the order based on the knowledge-based understanding of the parameters with the highest to lowest potential correlation with arsenic. The results itself validate the feasibility of ANFIS in identifying correlating variables. Thus, for ANFIS, only variables that are closely associated with dependent variables can be used as input variables for the model estimates.

The model with PS8 as inputs was found to be the best-fit model, with $R^2$ value reaching up to 0.998. This result indicates that the optimum variable group comprised the selected eight geochemical parameters, which reflects the geochemical processes related with arsenic enrichment. The corresponding geochemical processes impose the joint constraint on the arsenic concentration involving Fe/Mn redox processes, phosphate-competitive adsorption, pH-dependent desorption processes, and siderite precipitation.

MLR model

Eight-variable MLR model. Casewise diagnosis was initially employed to remove the outliers to form a good eight-variable MLR model, because the eight variables have been validated by ANFIS as the optimum variable group. The $R^2$ values increased evidently from 0.29 to 0.43 after casewise diagnosis, whereas the standard error of the estimates drastically declined from 124.97 to 41.12 (Table 3). These variations indicate that outliers remarkably inhibit the entire goodness of the model regression. Therefore, the derived model exhibited better performance after the outliers were removed.

We present the two eight-variable models before and after removing the outliers in Table 4 to illustrate the variations in the regression coefficients and demonstrate the influence of outliers on arsenic concentration. The results show that the standardized regression coefficients of HCO$_3^-$, dissolved P,
and FI exhibited a sign-reversing variation. Through case-wise diagnosis, we found that the selected outliers were responsible for the high concentration arsenic groundwaters (HCAG), which were precisely the outliers displayed in the binary correlation of the dissolved iron and arsenic (Fig. S1, part B; see the online supplementary materials). Therefore, the sign-reversal of the regression coefficients must have been caused by HCAG. This important information is expected to elucidate the geochemical effects resulting from the sign changes in the coefficients of $\text{HCO}_3^-$, dissolved P, and FI, which is discussed in subsequent sections.

**Optimum regression model.** After removal of the outliers, the statistical optimum regression model was obtained using stepwise regression.$^{[14,15]} \text{Fe}^{2+}$, $\text{Mn}^{2+}$, and $\text{SI}_{\text{ siderite}}$ were selected as independent variables in the resulting three-step

---

**Fig. 4.** Relationship of arsenic concentration calculated from ANFIS with observed values. (a), (b), (c), (d), and (e) represent PS8, PS5, MH8, MH5, and HH8, respectively.
Geochemical processes controlling arsenic enrichment in groundwater

Fig. 5. Comparison of calculated and observed arsenic concentrations. (a), (b), (c), and (d) represent PS8, PS5, MH8, and MH5, respectively. HH8 was not considered for comparison because of poor performance (color figure available online).

The results indicate a significant correlation between the independent variables and arsenic. The entire goodness of fit of the model did not remarkably change compared with the eight-variable models. However, slight variations were observed in $R^2$ and the standard error of the estimates (Tables 3 and 5).

Fe$^{2+}$, Mn$^{2+}$, and SI siderite were selected as independent variables for the statistical optimum regression model (Table 5), indicating that these three variables bear the closest associations with arsenic among the eight variables. The observed relationships of Fe$^{2+}$ and Mn$^{2+}$ with arsenic were consistent with the practical understanding of the mechanism of arsenic mobilization with reductive dissolution of the two elements in groundwater. Additionally, the selection of SI siderite into the optimum regression model may suggest that the close associations of dissolved iron with arsenic may have been caused by the constraints on arsenic by the dissolution equilibrium of siderite and the reductive dissolution of Fe(III).

Table 2. Statistical parameters used to assess the performance of the different ANFIS models ($R^2$ and RMSE).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>PS8</th>
<th>PS5</th>
<th>MH8</th>
<th>MH5</th>
<th>HH8</th>
<th>HH5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.997</td>
<td>0.988</td>
<td>0.978</td>
<td>0.746</td>
<td>0.442</td>
<td>—</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.008</td>
<td>0.029</td>
<td>0.022</td>
<td>0.083</td>
<td>0.119</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 3. Statistical parameters used to assess the regression performance of the eight-variable MLR models reported as correlation coefficient, $R^2$, and standard error of the estimates.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>$R^2$</th>
<th>Std. Error*</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 variables model with all samples as cases</td>
<td>0.54</td>
<td>0.29</td>
<td>124.97</td>
</tr>
<tr>
<td>8 variables model with outliers left out</td>
<td>0.66</td>
<td>0.43</td>
<td>41.12</td>
</tr>
</tbody>
</table>

*Std. Error denotes standard error of the estimates. The noted symbols are maintained consistently in the following tables.
The optimum regression model showed significant correlations of Fe²⁺ and Mn²+ with arsenic (P ≤ 0.01, Table 5), suggesting that the reductive dissolution of iron/manganese oxides is the dominant geochemical process for arsenic mobilization. The orderly selection of Fe²⁺ and Mn²+ is consistent with the general superiority of iron over manganese based on the acknowledged understanding of the reductive dissolution mechanism, i.e., the reductive dissolution of Fe(III) is more commonly reported to be responsible for arsenic mobilization than that of manganese.

S₁siderite was the third selected optimum variable in the optimum regression model (Table 5). The 0.01 level of significance of the standardized regression coefficient suggests that S₁siderite is a significant factor correlated with arsenic concentration (Table 5). In addition, we extracted the samples with S₁siderite less than zero to build an MLR model. The results show a significant positive relationship between the dissolved iron and arsenic (P = 0.00, Table 6). The R² value indicated that 78% of the arsenic variation could be explained by the dissolved iron, indicating a good correlation relationship (Table 6).

Table 4. Eight variable MLR models with standardized regression coefficients and significance of the t-test.

<table>
<thead>
<tr>
<th>Variables</th>
<th>8 variable model with all samples as cases</th>
<th>8 variable model with outliers left out</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Std. Coef.*</td>
<td>t</td>
</tr>
<tr>
<td>HCO₃⁻</td>
<td>0.15</td>
<td>1.22</td>
</tr>
<tr>
<td>NH₄⁺</td>
<td>0.13</td>
<td>0.99</td>
</tr>
<tr>
<td>P</td>
<td>-0.14</td>
<td>-1.16</td>
</tr>
<tr>
<td>Fe²⁺</td>
<td>0.13</td>
<td>0.90</td>
</tr>
<tr>
<td>Mn²⁺</td>
<td>0.26</td>
<td>2.42</td>
</tr>
<tr>
<td>FI</td>
<td>-0.03</td>
<td>-0.25</td>
</tr>
<tr>
<td>pH</td>
<td>0.20</td>
<td>1.79</td>
</tr>
<tr>
<td>S₁siderite</td>
<td>0.13</td>
<td>0.86</td>
</tr>
</tbody>
</table>

*Std. Coef. denotes standardized regression coefficient; P is the significance of statistical “t-test.” The noted symbols are maintained consistently in the following tables.

Discussion

Implications of Fe, Mn, and S₁siderite on the predominant geochemical process for arsenic enrichment

The origin of HCO₃⁻ in groundwater involves various geochemical processes, such as carbonate dissolution, silicate hydrolysis, and redox processes. The plus-to-minus sign variation of HCO₃⁻ in the eight-variable models after the removal of outliers indicated that HCO₃⁻ positively correlated with arsenic for HCAG. Such correlation suggests that HCO₃⁻ in these groundwaters more likely originated from the redox processes associated with arsenic mobilization. By contrast, HCO₃⁻ in low-concentration arsenic groundwater (LCAG) was mainly ascribed to other geochemical processes, such as the weathering of carbonate and silicate.

A high FI indicates that DOM was obtained from a microbial source. Thus, a good correlation of FI with arsenic suggests a connection between biogeochemical process and arsenic mobilization. The negative-to-positive sign variation in FI (Table 4) suggests that DOM is a covariate with arsenic concentration when HCAG is removed. This phenomenon implies that certain microbial activities or metabolisms, particularly those mediating reductive dissolution, could be inhibited in HCAG.

The sign variation in the phosphorus coefficient (Table 4) suggests that P is negatively correlated with arsenic for HCAG but is positively correlated with arsenic for LCAG.

Table 5. Statistical optimum regression model with variable selection by forward stepwise regression R² and the standard error.

<table>
<thead>
<tr>
<th>Model step</th>
<th>Variables selected for model</th>
<th>Std. Coef.</th>
<th>t</th>
<th>P</th>
<th>R</th>
<th>R²</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fe²⁺</td>
<td>0.53</td>
<td>5.83</td>
<td>0.00</td>
<td>0.53</td>
<td>0.28</td>
<td>44.42</td>
</tr>
<tr>
<td>2</td>
<td>Fe²⁺</td>
<td>0.51</td>
<td>5.90</td>
<td>0.00</td>
<td>0.59</td>
<td>0.35</td>
<td>42.37</td>
</tr>
<tr>
<td>3</td>
<td>Fe²⁺</td>
<td>0.27</td>
<td>3.12</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mn²⁺</td>
<td>0.29</td>
<td>2.49</td>
<td>0.01</td>
<td>0.63</td>
<td>0.40</td>
<td>40.94</td>
</tr>
<tr>
<td></td>
<td>Mn²⁺</td>
<td>0.27</td>
<td>3.27</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S₁siderite</td>
<td>0.31</td>
<td>2.68</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This phenomenon can be explained by the competitive adsorption effect, which is an established interpretation of the influence of phosphate on arsenic in Bangladesh aquifers. To further validate this effect on HCAG, we chose 170 µg/L as the boundary based on the results of casewise diagnosis, and samples with the arsenic concentration above this boundary (SCAB) were grouped to build stepwise MLR models. The model exactly selected P as the first optimum regression variable and derived a negative regression coefficient with a significant correlation with arsenic (P = 0.01, Table 7). Approximately 63% of change in arsenic could be explained by phosphorus, indicating a satisfactory goodness of fit (Table 7). These results strongly support the competitive adsorption effect of phosphate on arsenic for SCAB. The binary correlation between P and Ar are plotted in Figure 7A.

We found that pH was the second selected regression variable for SCAB (Table 7) when the scope of variable selection was broadened by adjusting the significance threshold for the t-test from 0.05 to 0.25. Moreover, pH exhibited a good covarying trend with arsenic (Fig. 7b). Furthermore, an unexpected negative correlation was observed between the dissolved iron and arsenic (Fig. 7c, R² = 0.15). This phenomenon suggests that the sorption effects from the phosphate-competitive adsorption and the pH elevation for HCAG are responsible for arsenic enrichment other than the reductive dissolution of iron oxides.

### Potential dynamics of reductive dissolution

The ANFIS and the three-variable optimum MLR models suggest that the reductive dissolution of Fe(III) favors arsenic enrichment. However, available information based on the contrasting models with and without outliers and the additional models through reselecting cases or adjusting the significance threshold reflected a great distinction in the relationships of the relevant geochemical parameters between LCAG and HCAG. This result suggests that the reductive dissolution of iron oxides may no longer be the dominant geochemical factor causing the arsenic enrichment of HCAG.

The results of the laboratory experiments demonstrated that reductive dissolution of iron oxides do not necessarily contribute to the elevation of arsenic concentration. On the contrary, the concentration of aqueous arsenic could be decreased because of the enhanced adsorption capability of the formed secondary minerals of iron oxides during microbe-mediated reductive dissolution of iron oxides. In a real aquifer, the majority of ions, such as cations, Fe²⁺, Mn²⁺, NH₄⁺, and anions, such as HCO₃⁻, would become highly concentrated with increasing aqueous arsenic accompanied with a series of redox transformations. Therefore, for the groundwaters with high arsenic concentration (i.e., >170 µg/L), reductive dissolution of iron oxides presumably approached a dynamic equilibrium. By contrast, the desorption effect would become more sensitive because of the limited amount of sorption sites on the solid surface as the ionic strength increases.

### Table 7. Stepwise regression model with SCAB as cases and statistical assessment indices

<table>
<thead>
<tr>
<th>Model step</th>
<th>Variables selected for model</th>
<th>Std. Coef. t</th>
<th>P</th>
<th>R</th>
<th>R²</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P</td>
<td>-0.79</td>
<td>-3.66</td>
<td>0.01</td>
<td>0.79</td>
<td>0.63</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>-0.65</td>
<td>-3.02</td>
<td>0.02</td>
<td>0.85</td>
<td>0.73</td>
</tr>
</tbody>
</table>

*The model given in Table 7 represents the stepwise regression model with SCAB as cases and a cutting off significance value equals to 0.05; we did not show the single model with significance cutoff value equals to 0.05, as the model with default cut off value of 0.05 is same with the results of step one of this model.*
The sign variations of $\text{HCO}_3^-$ and FI indicate the higher productivity of $\text{HCO}_3^-$ and lower microbial activity or metabolism capacity favoring the redox process for HCAG. This result suggests the ineffectiveness of further energetic reductive dissolution processes. At this point, the desorption effect from the competitive adsorption of phosphorous or pH elevation (Figs. 7a and b) could be the predominant controlling factor of arsenic concentration. Consequently, the evolutionary dominance of specific geochemical process in an aquifer may be inferred to control arsenic concentration.

Conclusions and implications

ANFIS and MRL were employed to develop multivariate models using 97 data sets from a routine high arsenic groundwater investigation at JHP. ANFIS models indicated that $\text{HCO}_3^-$, $\text{NH}_4^+$, $\text{P}$, $\text{Fe}^{2+}$, $\text{Mn}^{2+}$, FI, pH, and the siderite saturation compose the optimum variable group correlated with arsenic concentration. This result suggests that the geochemical processes involving reductive dissolution of iron/manganese oxides, phosphate-competitive adsorption, pH-dependent sorption, and siderite precipitation may pose an integrated constraint on arsenic concentration. The analysis of the different MLR models revealed that for LCAG, iron reductive dissolution was primarily responsible for arsenic mobilization. By contrast, for HCAG, reductive dissolution of iron oxides presumably approached a dynamic equilibrium. However, the desorption effects from the phosphate-competitive adsorption and pH elevation exhibited superior arsenic enrichment over other geochemical processes as the groundwater chemistry evolved. The inhibition effect of siderite precipitation on arsenic mobilization may exist for groundwaters with high $\text{SI}_{\text{siderite}}$.

This study offered a statistical method of interpreting various geochemical processes on arsenic enrichment using various models. The results not only correspond well with the knowledge-based understanding of arsenic geochemistry but also suggest an evolutionary dominance of specific geochemical processes in controlling the arsenic
concentration, which presented a heterogeneous distribution in groundwaters of an aquifer. Thus, the selection of a proper investigation scale would be a solution for mitigating the interference of heterogeneous geochemical effects.

Acknowledgments

This research work was financially supported jointly by the National Natural Science Foundation of China (41002088, 41302187), Institute of Hydrogeology and Environmental Geology, Chinese Academy of Geological Sciences (SK201303) and Chinese Geological Survey (1212011302700).

References


