Use of redundant data to reduce estimation errors in geochemical speciation

F. De Gaspari[a,b,*], M.W. Saaltink[a], J. Carrera[b], L.J. Slooten[b,c]

[a] *GHS, Department of Geotechnical Engineering and Geosciences, Universitat Politecnica de Catalunya, UPC-BarcelonaTech, c/Jordi Girona 1-3, 08034 Barcelona, Spain*
[b] *GHS, Institute of Enviromental Assessment and Water Research (IDÆA), CSIC, c/Jordi Girona 18, 08034 Barcelona, Spain*
[c] *Alten Nederland, Rivium 1e straat 85, 2909 LE Capelle a/d IJssel, Nederland*

Abstract

Speciation is the process of evaluating the concentrations of all the species in a chemical system from equilibrium conditions and measured data such as total concentrations of components, electrical conductivity, pH, redox potential, gas partial pressure. It is essential for analyzing geochemical data and defining the chemical composition of waters for geochemical modeling problems like evaluating the chemical composition of evaporating, diluting, mixing waters or reactive transport. We present an algorithm that reduces estimation errors in chemical speciation calculations by means of the use of redundant data. Redundant data are measurements and assumptions that exceed the minimum data set required, and therefore are not strictly necessary, to speciate a water sample. The proposed method was compared with the classical speciation algorithm on two synthetic examples. Our re-

*Corresponding author
Email addresses: francesca.de.gaspari@upc.edu (F. De Gaspari), maarten.saaltink@upc.edu (M.W. Saaltink), jesus.carrera.ramirez@gmail.com (J. Carrera), luitjan.slooten@gmail.com (L.J. Slooten)

sults show that using redundant data improves speciation results reducing the estimation error between computations and measurements. Moreover, the larger the amount of redundant data, the better in terms of errors of the estimated concentrations.
Keywords:
Redundant data, Geochemical modeling, Speciation, Optimization problem

1. Introduction

Geochemical modeling is important in Earth Sciences. In particular, it is required to assess problems ranging from weathering to the characterization of the chemical composition of water and processes that could influence its quality (Appelo and Postma, 2010; Bethke, 2008). Geochemical speciation is a key step of geochemical modelling that consists of evaluating concentrations of all the species in a chemical system from measured data (e.g., total concentrations of components, pH, alkalinity, gas partial pressures, electrical conductivity, redox potential) and equilibrium constraints. For this reason, it is often termed thermodynamic speciation.

Speciation requires the solution of a non-linear system of equations and a lot of research has been focused on numerical issues that might arise when solving these equations. Several methods have been proposed to solve chemical equilibrium in a robust way in order to guarantee the convergence (Paz-García et al., 2013; Carrayrou et al., 2002; Brassard and Bodurtha, 2000) and many codes have also been released to deal with geochemical speciation calculations: GEMS3K (Kulik et al., 2013), Visual MINTEQ (Gustafsson, 2011), CHEPROO (Bea et al., 2009), ORCHESTRA (Meeussen, 2003),

MIN3P (Mayer et al., 2002), PHREEQC (Parkhurst et al., 1999) and its interactive version, PHREEQCi (Charlton et al., 1997), EQ3NR (Wolery, 1983, 1992) and WATEQ4F (Ball and Nordstrom, 1991).

Speciation calculations are subject to implicit sources of uncertainty which can derive from uncertainty in thermodynamic data, such as equilibrium constant values, or from errors in chemical analyses (i.e., analytical errors). These types of random errors can be referred to as "aleatory uncertainty". Misjudgment in the definition of the chemical system, such as failure to account for some reactions or discarding others, can also lead to errors in speciation. These can be defined as "epistemic uncertainty". They arise from an incomplete or inadequate characterization of the system (Gupta et al., 2012), such as assuming the neutrality of a solution when it is not electrically balanced, or imposing equilibrium with phases that are not. The effect of errors propagation in geochemical calculations has been extensively studied. In particular, the effect of aleatory errors has been investigated by Weber et al. (2006); Denison and Garnier-Laplace (2005); Ödegaard-Jensen et al. (2004); Nitzsche et al. (2000); Cabaniss (1999, 1997); Criscenti et al. (1996); Merino (1979), while Smith et al. (1999) examined the connection between aleatory and epistemic errors. Although the origin and propagation effects of both types of errors are different, they can be treated in the same way through probability density functions, e.g., by means of mean and standard deviation values.

All these studies use a fixed number of data to solve the speciation. Geochemical speciation, in fact, requires a fixed minimum number of data, including equilibrium assumptions, equal to the number of independent variables

of the system (i.e., number of species). For example, a carbonate system is characterized by four degrees of freedom (see Section 2.1). Therefore, four data (e.g., total concentrations of inorganic carbon and calcium, pH) or hypotheses (e.g., water activity equal to 1) are needed. However, extra data might be available (e.g., alkalinity, electrical conductivity or redox potential) or extra assumptions about the system might be made (e.g., equilibrium with calcite or $CO_{2(g)}$ in equilibrium with the atmosphere). Chemical analyses of waters, for example, often provide extra data and also the analytical errors associated to each of them.

We term these extra data as redundant and we claim that speciation calculations can benefit from their use, while aknowledging analytical errors.

The aim of this paper is to present an algorithm to include redundant data in speciation calculations and to prove that their use can improve the results by reducing estimation errors. We also claim that increasing the number of redundant data helps decreasing the estimation errors even further.

2. Methodology

We start by analyzing a speciation example to clarify the differences between the traditional and the proposed method. This allows us to formalize the problem statement and to propose a solution algorithm.

2.1. Speciation of a carbonate system

We consider the problem of calculating the concentrations of dissolved species in a carbonate system. This system has received extensive attention from the scientific community, e.g. to study seawater intrusion in carbonate coastal aquifers (Werner et al., 2013; Bear, 1999; Back et al., 1979, amongst

many others), including geochemical processes occurring in the mixing zone between freshwater and saltwater (Sanz et al., 2011; De Simoni et al., 2007; Rezaei et al., 2005), and to analyze the feasibility of CO_2 sequestration in deep aquifers (Saaltink et al., 2013; Duan and Li, 2008; Xu et al., 2006).

The most simple chemical system consists of 9 species ($N_s = 9$) and the following 5 equilibrium reactions ($N_{re} = 5$):

$$\begin{aligned}
&OH^- + H^+ \rightleftharpoons H_2O && \log K_1 = 13.995 \\
&CO_3^{2-} + H^+ \rightleftharpoons HCO_3^- && \log K_2 = 10.328 \\
&CO_{2(aq)} + H_2O \rightleftharpoons HCO_3^- + H^+ && \log K_3 = -6.344 \\
&CO_{2(g)} + H_2O \rightleftharpoons HCO_3^- + H^+ && \log K_4 = -7.813 \\
&CaCO_{3(s)} + H^+ \rightleftharpoons Ca^{2+} + HCO_3^- && \log K_5 = 1.848
\end{aligned} \tag{1}$$

The number of degrees of freedom of this system is $N_s - N_{re} = 4$. This means that 4 data or assumptions are needed to solve the speciation problem. Speciation codes normally use this criterion. Optionally species with constant activity can be decoupled and eliminated, e.g., water if the system is sufficiently diluted ($a_{H_2O} = 1$) or proton if the pH is fixed ($a_{H^+} = 10^{-pH}$), to reduce the number of unknowns. Numerous methods have been proposed to eliminate constant activity species in reactive transport calculations (Kräutle, 2011; De Simoni et al., 2005; Kräutle and Knabner, 2005; Molins et al., 2004; Saaltink et al., 1998). Regardless of the decision to eliminate them, we refer generically to these methods as the traditional speciation methods, as they should all yield the same results.

Being the degrees of freedom for system (1) equal to 4, the concentrations

of all species can be calculated from four known data: total concentration of calcium, alkalinity, activity of water and pH for example

$$\begin{cases} Ca_{tot} : [Ca^{2+}] - x_1 = 0 \\ Alkalinity : [HCO_3^-] + 2[CO_3^{2-}] + [OH^-] - [H^+] - x_2 = 0 \\ Water\ Activity : \mathrm{a}_{H_2O} - x_3 = 0 \\ pH : -\log \mathrm{a}_{H^+} + x_4 = 0 \end{cases} \tag{2}$$

where [ ] represents molal concentration (mol/kgw). x_1, x_2 and x_4 are actual measurements representing Ca_{tot}, $Alkalinity$ and pH, while x_3 is the value of water activity fixed to 1. We term these kind of equations "data equations". These must be solved together with the mass action laws deriving from system (1)

$$\mathbf{f}_{MAL} = \mathbf{S}_e \log \mathbf{a} - \log \mathbf{k} = 0 \tag{3}$$

where $\mathbf{a}$ is a vector containing the activities of the N_s species, $\mathbf{S}_e$ is a matrix $(N_{re} \times N_s)$ with the stoichiometric coefficients of the equilibrium reactions and $\mathbf{k}$ is a vector (N_{re}) of equilibrium constants.

Generalizing the traditional speciation method we can say that $N_1 = N_s - N_{re}$ data equations need to be solved together with $N_2 = N_{re}$ mass action laws, $\mathbf{f}_{MAL}$:

$$\begin{cases} \mathbf{g}(\mathbf{c}) - \mathbf{x} = 0 \\ \mathbf{f}_{MAL}(\mathbf{c}) = 0 \end{cases} \tag{4}$$

where $\mathbf{c}$ is the vector of concentrations of the N_s species, $\mathbf{x}$ a vector of N_1 data and $\mathbf{g}(\mathbf{c})$ defines operations to be applied to $\mathbf{c}$ in order to compute what

is measured (e.g., linear combinations of species concentrations to obtain measured components, or $-\log(\gamma_{H^+} \cdot [H^+])$ to obtain pH, where γ_{H^+} is the proton activity coefficient). Typically data equations contain balances of total concentrations, electrical charge, alkalinity, total dissolved inorganic carbon (TIC), pH values, redox potential or electrical conductivity.

The traditional algorithm to speciate consists of five steps: (1) dividing the species in two sets of $N_1 = N_s - N_{re}$ primary and $N_2 = N_{re}$ secondary species (Steefel and Yabusaki, 2000) with concentrations $\mathbf{c}_1$ and $\mathbf{c}_2$, respectively; (2) guess an initial value of primary concentrations; (3) use $\mathbf{f}_{MAL}$ to calculate $\mathbf{c}_2 = f(\mathbf{c}_1)$; (4) use data $\mathbf{x}$ to solve $g(\mathbf{c}_1, \mathbf{c}_2) - \mathbf{x} = 0$ for $\mathbf{c}_1$, (5) repeat steps (3) and (4) until convergence.

This work is focused on cases in which the number of available data is larger than N_1. In this case, the resulting data equations cannot be solved exactly. Instead, they need to aknowledge measurement errors.

For example, if measurements of total dissolved inorganic carbon (TIC) and pressure of gas ($P_{CO_{2(g)}}$) were available and we wanted to apply zero charge balance and equilibrium with calcite as well, the data equations could be rewritten as

$$\begin{cases} Ca_{tot} : [Ca^{2+}] - x_1 = \varepsilon_1 \\ Alkalinity : [HCO_3^-] + 2[CO_3^{2-}] + [OH^-] - [H^+] - x_2 = \varepsilon_2 \\ Water\ Activity : \mathrm{a}_{H_2O} - x_3 = \varepsilon_3 \\ pH : -\log \mathrm{a}_{H^+} + x_4 = \varepsilon_4 \\ TIC : [CO_{2(aq)}] + [HCO_3^-] + [CO_3^{2-}] - x_5 = \varepsilon_5 \\ P_{CO_{2(g)}} : \log \mathrm{a}_{CO_{2(g)}} - x_6 = \varepsilon_6 \\ Charge\ Balance : [H^+] + 2[Ca^{2+}] - [OH^-] - [HCO_3^-] - 2[CO_3^{2-}] = \varepsilon_7 \\ Calcite\ Eq. : \log \mathrm{a}_{\mathrm{Ca^{2+}}} + \log \mathrm{a}_{\mathrm{HCO_3^-}} - \log \mathrm{a}_{\mathrm{H^+}} - \log K_5 = \varepsilon_8 \end{cases} \tag{5}$$

where x_5 is the measured TIC, x_6 is $\log(P_{CO_{2(g)}})$ and x_7 and x_8 are equal to zero because of the zero charge balance and equilibrium constraints (x_7 corresponds to the saturation index of calcite, null at equilibrium). ε_i, $i = 1, ..., 8$, represent measurement errors that need to be taken in account since the system to be solved has become overdetermined (i.e., the number of data is larger than N_1). The data set (5) presents 4 redundant data.

The algorithm to solve data equations (5) together with mass action laws (3) to speciate is explained in the following section.

2.2. Speciation with redundant data: Problem statement

If redundant informations are used to solve a speciation problem, system (4) can be re-defined as follows

$$\begin{cases} \mathbf{g}(\mathbf{c}) - \mathbf{x} = \boldsymbol{\varepsilon} \\ \mathbf{f}_{MAL}(\mathbf{c}) = 0 \end{cases} \tag{6}$$

The differences of system (6) from the traditional speciation problem defined in (4) are the dimension of $\mathbf{g}$ and $\mathbf{x}$ ($dim(\mathbf{g}) = dim(\mathbf{x}) = N_d > N_1$) and errors in measurements $\boldsymbol{\varepsilon}$ which are included. $\boldsymbol{\varepsilon}$ can incorporate analytical errors in data, such as in data 1 to 4 in system (5), and uncertainty about the correct model, such as charge balance and equilibrium with calcite assumptions in system (5).

When solving speciation problems, it is common to use data equations which are either linear combinations of concentrations (e.g., TIC, alkalinity) or linear combinations of log-activities (e.g., equilibrium with minerals or gases). Moreover, the errors ($\boldsymbol{\varepsilon}$) of both types of data equations can have a normal or log-normal distribution. Therefore, the expressions of $\mathbf{g}(\mathbf{c})$ must be defined and calculated accordingly to the types of data equations (see Appendix A for details).

System (6) is overdetermined, therefore a non-linear least square fitting is required to minimize $\boldsymbol{\varepsilon}$, as described below.

2.3. Algorithm

We want to find the solution of (6) that minimizes the sum S of the weighted squares of the difference between measured and calculated data, defined as

$$S = \boldsymbol{\varepsilon}^t \mathbf{V}^{-1} \boldsymbol{\varepsilon} \tag{7}$$

where $\mathbf{V}$ is the covariance matrix ($N_d \times N_d$) of measurement errors. Without loss of generality, we will assume errors to be not correlated, so that $\mathbf{V}$ is a diagonal matrix, containing the variance of each i-th measurement, $\sigma^2_{e,i}$.

The condition leading to the minimum value of S is that its derivative with respect to the unknowns, $\ln \mathbf{c}_1$, is zero:

$$\frac{\partial S}{\partial \ln \mathbf{c}_1} = 2\boldsymbol{\varepsilon}^t \mathbf{V}^{-1} \frac{\partial \mathbf{g}}{\partial \ln \mathbf{c}_1} = 2\boldsymbol{\varepsilon}^t \mathbf{V}^{-1} \mathbf{J} = \mathbf{0} \tag{8}$$

where $\mathbf{J}$ is the jacobian matrix containing the derivatives of $\mathbf{g}(\mathbf{c})$ with respect to the unknowns, whose expression is derived and explained in Appendix B. We decided to work with $\ln \mathbf{c}_1$ as variable but it would be equally possible to express all the equations as function of $\mathbf{c}_1$.

Approximating the function $\boldsymbol{\varepsilon}$ linearly between two sequential iterations k and $k+1$

$$\boldsymbol{\varepsilon}^{k+1} = \boldsymbol{\varepsilon}^k + \mathbf{J}^k \Delta \ln \mathbf{c}_1^k \tag{9}$$

with

$$\Delta \ln \mathbf{c}_1^k = \ln \mathbf{c}_1^{k+1} - \ln \mathbf{c}_1^k \tag{10}$$

and substituting it in (8), the solution for a given iteration k is

$$\mathbf{J}^t \mathbf{V}^{-1} \mathbf{J} \Delta \ln \mathbf{c}_1 = -\mathbf{J}^t \mathbf{V}^{-1} \boldsymbol{\varepsilon} \tag{11}$$

After convergence, the covariance matrices of the estimation errors associated to $\ln \mathbf{c}_1$ ($\boldsymbol{\Sigma}_1$) and $\ln \mathbf{c}_2$ ($\boldsymbol{\Sigma}_2$) can be calculated. This can be useful to analyze the quality of the estimation (see section 2.4). $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, in fact, provide the uncertainty associated to the estimation of $\ln \mathbf{c}_1$ and $\ln \mathbf{c}_2$. $\boldsymbol{\Sigma}_1$ ($N_1 \times N_1$) can be calculated by means of the "real" covariance matrix of the data, $\mathbf{C}_d$ ($N_d \times N_d$):

$$\boldsymbol{\Sigma}_1 = Cov(\ln \mathbf{c}_1) = (\mathbf{J}^t \mathbf{C}_d^{-1} \mathbf{J})^{-1} \tag{12}$$

168 However, in reality $\mathbf{C}_d$ is not known. Therefore, it is necessary to make
169 an hypothesis about its structure. A reasonable assumption defines $\mathbf{C}_d$ as
170 proportional to S and to the covariance of measurement errors, $\mathbf{V}$:

$$\mathbf{C}_d = \sigma^2 \mathbf{V} \tag{13}$$

171 where $\sigma^2 = S/N_d$ (S was defined in equation 7). Substituting (13) into
172 (12) we obtain

$$\boldsymbol{\Sigma}_1 = \sigma^2 (\mathbf{J}^t \mathbf{V}^{-1} \mathbf{J})^{-1} \tag{14}$$

173 $\boldsymbol{\Sigma}_2$ ($N_2 \times N_2$) can be calculated by taking into account the dependence
174 of $\mathbf{c}_2$ on $\mathbf{c}_1$

$$\boldsymbol{\Sigma}_2 = Cov(\ln \mathbf{c}_2) = \left(\frac{\partial \ln \mathbf{c}_2}{\partial \ln \mathbf{c}_1} \right) \boldsymbol{\Sigma}_1 \left(\frac{\partial \ln \mathbf{c}_2}{\partial \ln \mathbf{c}_1} \right)^t \tag{15}$$

175 Details on the calculation of $(\partial \ln \mathbf{c}_2 / \partial \ln \mathbf{c}_1)$ are explained in Appendix
176 B.

177 The steps of the proposed algorithm can be outlined as follows:

178 1. Set $\mathbf{x}$ and matrices of $\mathbf{g}(\mathbf{c})$ (see Appendix A)

179 2. Guess initial value of $\mathbf{c}_1^0$

180 3. Given $\mathbf{c}_1^k$, calculate secondary concentrations $\mathbf{c}_2^k = f(\mathbf{c}_1^k)$ and $\partial \mathbf{c}_2^k / \partial \mathbf{c}_1^k$
181 from $\mathbf{f}_{MAL} = 0$ (see Appendix B)

4. Calculate $\boldsymbol{\varepsilon}^k$, Jacobian matrix $\mathbf{J}^k$, and RHS and LHS of system (11). Solve system (11) and evaluate $\Delta \ln \mathbf{c}_1^k$

5. Update the solution $\ln \mathbf{c}_1^{k+1} = \ln \mathbf{c}_1^k + \Delta \ln \mathbf{c}_1^k$

6. Set k=k+1 and repeat steps 3. to 5. until convergence

7. After convergence, calculate $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$

As convergence criteria, we check the maximum relative error between two sequential iterations and the residual (RHS in system 11) at every iterations. The iterative process is stopped when both quantities are smaller than threshold values defined by the user.

2.4. Testing approach

The algorithm was tested by means of two synthetic examples: first a single reaction representing gypsum equilibrium at a temperature of 25 ^{o}C in ideal conditions ($\mathbf{a} = \mathbf{c}$)

$$CaSO_{4(s)} \rightleftharpoons Ca^{2+} + SO_4^{2-} \qquad \log K = -4.482 \tag{16}$$

We chose arbitrarly a known solution of (16) and 5 possible measurements with errors of $\log[Ca^{2+}]$ and $\log[SO_4^{2-}]$, decimal logarithm of calcium and sulfate concentrations, respectively. Figure 1 shows the equilibrium line and the exact solution of the speciation together with the 5 measurement points that were used in this case. Both the traditional and the proposed methods were applied to solve the speciation. Since the system is characterized by one degree of freedom, i.e. one datum is necessary to solve the speciation, first the traditional method was employed using only $\log[Ca^{2+}]$ data, and

later the proposed method was employed using both $\log[Ca^{2+}]$ and $\log[SO_4^{2-}]$ measurements. The results were first compared in terms of logaritmic mean squared error, MSE_{log}:

$$MSE_{log} = \frac{1}{N}\frac{1}{N_s}\sum_{i=1}^{N}\sum_{j=1}^{N_s}\left[\log\left(\frac{c_{ij}}{c_j^*}\right)\right]^2 \tag{17}$$

where N is the number of measurements (5), N_s is the number of species (2), c_{ij} is the calculated concentration of the i-th measurement and j-th species, and c_j^* is the exact value of the j-th species. Comparing the two methods in terms of MSE_{log} was possible because the exact solution in this case is known. However, when a speciation is calculated using real data the exact solution is not known a priori. Therefore, we computed the covariance matrices of estimation errors $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ defined in (14) and (15), respectively, and compared the estimation errors of the traditional and the proposed methods also in terms of the variable Var^*:

$$Var^* = \frac{1}{N}\frac{1}{N_s}\sum_{i=1}^{N}\left[\sum_{l=1}^{N_1}\left(\Sigma_{1,ll}\right)_i + \sum_{m=1}^{N_2}\left(\Sigma_{2,mm}\right)_i\right] \tag{18}$$

Var^* represents a mean value of the estimation errors variance of all the species. From (14) it can be noticed that $\mathbf{\Sigma}_1$ depends on the error $\boldsymbol{\varepsilon}$ through the variable σ^2. However, for the traditional speciation method $\boldsymbol{\varepsilon} = 0$, hence $S = 0 \Rightarrow \sigma^2 = 0$ (see system 4). This makes (14) not suitable for the comparison. Therefore, we fixed $\sigma^2 = 1$ for the two methods. This way, we still guarantee that the real covariance matrix of data, $\mathbf{C}_d$ (see equation 13), is the same for both methods.

An input variable of the method is the variance of measurement errors

associated to the data, σ_e^2 (see equation 7). In this example we used the same value for both $\log[Ca^{2+}]$ and $\log[SO_4^{2-}]$ data ($\sigma_e = 0.17$), so that they have the same weight. In order to make the results of this example more general we tested again the two methods for 1500 measurements of $\log[Ca^{2+}]$ and $\log[SO_4^{2-}]$ with $\sigma_e = 0.17$. The exact solution was therefore perturbed 1500 times by means of a lognormal distribution with standard deviation $\sigma_g = 0.17$ to obtain the measurement points shown in figure 2. The use of a lognormal distribution allowed us to avoid possible negative values in concentrations in the measurements generation process.

The advantage of this simple problem is that it presents an analytical solution. In the case of using only $\log[Ca^{2+}]$ data, the variance of $\log[Ca^{2+}]$ error will be σ_e^2 because $\log[Ca^{2+}]$ will remain unchanged. The same error will transfer to $\log[SO_4^{2-}] = \log K$ - $\log[Ca^{2+}]$. Therefore, the expected value of Var^* is σ_e^2. If both $\log[Ca^{2+}]$ and $\log[SO_4^{2-}]$ measurements are used, it is easy to check that minimizing $\varepsilon^2_{Ca^{2+}} + \varepsilon^2_{SO_4^{2-}}$, where $\varepsilon_i = \log c_i - x_i$ ($i = Ca^{2+}, SO_4^{2-}$), subject to $\log[Ca^{2+}]+\log[SO_4^{2-}] = \log K$, leads to $\log[Ca^{2+}] = (\log K + x_{Ca^{2+}} - x_{SO_4^{2-}})/2$. Thus, the variance of both estimates is $\sigma_e^2/2$, which is the expected value of Var^*. This result will serve to test our approach and to illustrate the advantage of using redundant data.

As second example we chose the carbonate system defined in (1). The extended Debye-Hückell expression for activity coefficients was used in this example (Helgeson and Kirkham, 1974). As explained in section (2.1) the system is characterized by four degrees of freedom. We used as exact solution a water in equilibrium with calcite, with partial pressure of $CO_{2(g)}$ equal to $10^{-3.5}$, $\mathrm{a}_{H_2O} = 1$ and electrically balanced. Its chemical composition is shown

in Table 1.

First we compared the traditional and the proposed methods to verify the accuracy of the two algorithms in terms of speciation results. For this purpose, six data were extracted from the exact solution to be used in the speciation calculations (Table 2). The traditional speciation method was employed using four data equations: $alkalinity$, Ca_{tot}, a_{H^+} and $a_{H_2O} = 1$. Note that the proton activity presents a measurement error, therefore cannot be eliminated. We will refer to this case as 'solution 1'. Afterwards we tested the proposed method adding gradually redundant data to the previous three: TIC ('solution 2'), activity of the gas, $a_{CO_{2(g)}}$ ('solution 3') and equilibrium with calcite condition ('solution 4'). Afterwards we compared the solution obtained with the traditional speciation method to the solutions obtained using an increasing number of redundant data: from 1 in solution 2 to 3 in solution 4. As for the gypsum example, we perturbed the data to generate 1500 possible measured values and then we compared the speciation results in terms of MSE_{log} and Var^*, defined in (17) and (18), respectively.

The measured values were generated perturbing the logarithm of the exact value (μ) with a standard deviation, σ_g, of 0.17 by means of a log-normal distribution. The condition of calcite equilibrium was not perturbed ($\sigma_g = 0$), since zero is the reference value of the saturation index for minerals in equilibrium. The values of σ_g were used also to define the uncertainty associated to every datum presenting analytical errors ($\sigma_e = 0.17$, see Table 2).

In reality, however, it is difficult to know the correct value of uncertainty for each type of measurement. To analyze the effect of an incorrect measurement error we performed a second group of simulations in which we changed

the σ_e values of all data one at a time and calculated MSE_{log} and Var^* as function of the standard deviation of measurement errors of every constraint. We chose two values of σ_e: the first larger than the one used to generate the perturbed measurements ($\sigma_e = 0.35 > \sigma_g$), to simulate a higher uncertainty associated to the data, and the second smaller ($\sigma_e = 0.09 < \sigma_g$), to simulate more certain data values.

3. Results

3.1. Gypsum example

Figure 3 shows the results of the traditional speciation method, i.e. using only $\log[Ca^{2+}]$ data. It can be noticed that the five points moved on the equilibrium line, since the equilibrium with gypsum was imposed as certain condition, along a line parallel to the y-axis which represents the imposed calcium concentration data. In this case MSE_{log}=0.28 and Var^*=0.154. Note that Var^* coincides with its expected value, σ_e^2, once ln is converted to $\log_{10}$. Afterwards, the proposed method was tested, i.e., using both $\log[Ca^{2+}]$ and $\log[SO_4^{2-}]$ data. The results are shown in figure 4. It can be observed that while some of the points moved further from the exact solution with respect to the classical speciation results (white and black triangles), the others moved closer to the exact solution. However, for all the points the proposed algorithm minimizes the distance between measured and calculated data. The calculated mean squared error in this case was 0.23, smaller than 0.28 obtained with the traditional method. Moreover, Var^*=0.077, which coincides again with its expected value, $\sigma_e^2/2$.

The same methodology was employed to compare the two methods for

the 1500 measurement points of figure 2 and the resulting mean squared error decreased from 0.029 for the traditional method to 0.016 for the proposed methods. The value of Var^* also decreased by half, confirming its expected value: from 0.156 to 0.078 for traditional and proposed methods, respectively.

3.2. Carbonate example

The results of the comparison between the two methods in terms of MSE_{log} are shown in figure 5. It can be noticed that the value of the mean squared error for the traditional method (solution 1) is barely larger than 0.05, while it is smaller for the solutions using redundant data (solutions 2, 3 and 4). Moreover, increasing the number of redundant data used in the speciation contributes to reduce more the MSE_{log} value: it decreases from 0.04 using only one redundant data (solution 2) to 0.016 using 3 redundant data (solution 4).

Figure 6 shows the effect of changing the standard deviation associated to measurements (σ_e). Obviously its value does not affect solution 1, which is the result of a traditional speciation calculation. Neither does the value of σ_e for $a_{CO_{2(g)}}$ have an effect on solution 2 (figure 6e) because this solution does not use $a_{CO_{2(g)}}$ data. Nor does the σ_e for the assumption of equilibrium with calcite have an effect on solutions 2 and 3 (figure 6f), for the same reason. In general one can observe that the use of an incorrect standard deviation can worsen the solution with respect to the one obtained with the correct standard deviations. However, the quality of the solution in terms of MSE_{log} improves using redundant data with respect to the traditional speciation, despite a wrong choice about σ_e value. Decreasing the uncertainty relative to the equilibrium with calcite assumption (figure 6f) improves the

solution even with respect to the one obtained with correct σ_e values. This is because the correct value of σ_g for this datum is 0, not 0.17 (see Table 2). The effect of σ_e relative to alkalinity and TIC (figure 6c, d) are very similar as their values are very close, due to the fact that in this pH range the concentration of HCO_3^- is predominant with respect to carbonate species or OH^- concentrations. Finally, it seems that changing the uncertainty relative to Ca_{tot} (figure 6a) does not affect the solution. Nevertheless, when a large number of redundant data is used, such as in solution 4, the standard deviation seems to have a minor effect on the estimation error, MSE_{log}.

The effect of an incorrect value of σ_e on Var^* was also analyzed. Only the results for a_{H^+} and alkalinity are reported in figure 7 as the most representative. It can be noticed that the error variance can be big for the traditional speciation method (solution 1), while it slightly decreases when redundant data are used (solutions 2, 3 and 4). Moreover, the more redundant data are used, the more the variance of estimation error decreases, converging to a value corresponding to the true standard deviation of measurement errors (σ_e=0.17).

4. Conclusions

We proposed a speciation algorithm that uses redundant data and aknowledges measurement errors, on the assumption that redundant data will reduce estimation errors in geochemical calculations.

We compared the proposed method with the traditional speciation method in terms of logaritmic mean squared error, MSE_{log} and mean value of estimation error variance, Var^*. We tested both algorithms by means of two

synthetic examples. Both MSE_{log} and Var^* using redundant data are consistently smaller than in the traditional method.

The effect of measurement errors was examined in a carbonate system example. The algorithm is sensitive to the variance of measurement errors. Also, a wrong value of the standard deviation can worsen the results with respect to the ones obtained with the correct standard deviation. However, the effect of its value depends on the type of data associated to it. A wrong error associated to measurements can still improve the results in terms of mean squared error and variance of estimation error with respect to a traditional speciation method, especially when a large number of redundant data are used.

Therefore we argue that the proposed method can improve the quality of the speciation results, reducing estimation errors.

Appendix A. Error definitions

The errors $\boldsymbol{\varepsilon}$ allowed in the proposed method can be additive or multiplicative. Additive errors should lead to gaussian distributions, whereas multiplicative to lognormal distributions. Depending on the type of error, the function $\mathbf{g}(\mathbf{c})$ and the data $\mathbf{x}$ must be defined accordingly: aritmetic or logaritmic for additive and multiplicative errors, respectively.

Data used in speciation calculations are typically combinations of concentrations or activity values. The former, which we name balance equations, are linear combinations of concentrations representing total concentrations, alkalinity, charge balance or TIC values. The latter are usually employed to fix pH values or to impose equilibrium with minerals or gases.

Distinguishing between these two types of data equations we can define:

$$\boldsymbol{\varepsilon} = \mathbf{Bc} - \mathbf{x} \tag{19}$$

for the balance equations and

$$\varepsilon_i = \prod_{n=1}^{N_s} {a_n}^{L_{in}} - x_i \tag{20}$$

for each $i-th$ activity combination, respectively. $\mathbf{B}$ is a matrix of dimension $(N_b \times N_s)$ and N_b is the number of balance equations. $\mathbf{B}$ contains different coefficients depending on the type of balance equation: ionic charge for charge neutrality, coefficients defining alkalinity or TIC, or component matrix elements for total concentration. $\mathbf{L}$ is a matrix of dimension $(N_a \times N_s)$ containing the coefficients of the activity for every species involved in the combination. N_a is the number of activity conditions imposed.

If we want to use a log-distribution instead of a normal distribution of errors, one should use:

$$\boldsymbol{\varepsilon} = \ln(\mathbf{Bc}) - \ln \mathbf{x} \tag{21}$$

for the balance equations and

$$\boldsymbol{\varepsilon} = \mathbf{L} \ln \mathbf{a} - \ln \mathbf{x} \tag{22}$$

for the activity combinations, respectively.

Appendix B. Jacobian calculation

The jacobian, containining the derivatives of $\boldsymbol{\varepsilon}$ with respect to the state variables $\ln \mathbf{c}_1$ at every step of the iterative method, can be calculated as

$$\begin{aligned} \frac{\partial \varepsilon_i}{\partial \ln c_{1,j}} &= \frac{\partial \varepsilon_i}{\partial c_{1,j}} \cdot c_{1,j} \\ &= \mathrm{B}_{1,ij} \cdot c_{1,j} + \sum_{l=1}^{N_2} \mathrm{B}_{2,il} \cdot \frac{\partial c_{2,l}}{\partial c_{1,j}} \cdot c_{1,j} \\ & i = 1, \ldots, N_b \\ & j = 1, \ldots, N_1 \end{aligned} \tag{23}$$

from the definition (19) whereas by means of definition (21) results

$$\begin{aligned} \frac{\partial \varepsilon_i}{\partial \ln c_{1,j}} &= \frac{\partial \ln z_i}{\partial \ln c_{1,j}} \\ &= \frac{1}{z_i} \cdot \frac{\partial z_i}{\partial \ln c_{1,j}} \\ &= \frac{1}{z_i} \cdot \left(\mathrm{B}_{1,ij} \cdot c_{1,j} + \sum_{l=1}^{N_2} \mathrm{B}_{2,il} \cdot \frac{\partial c_{2,l}}{\partial c_{1,j}} \cdot c_{1,j} \right) \\ & i = 1, \ldots, N_b \\ & j = 1, \ldots, N_1 \end{aligned} \tag{24}$$

being

$$z_i = \left(\sum_{m=1}^{N_1} \mathrm{B}_{1,im} \cdot c_{1,m} + \sum_{l=1}^{N_2} \mathrm{B}_{2,il} \cdot \frac{\partial c_{2,l}}{\partial c_{1,j}} \cdot c_{1,j} \right) \tag{25}$$

Matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ are the parts of matrix $\mathbf{B}$ relative to primary and secondary species, respectively, and the derivatives of secondary concentrations with respect to primary concentrations can be calculated considering

that at every step of the iterative method the total derivative of $\mathbf{f}_{MAL}$ with respect to primary species concentrations is null:

$$\frac{d\mathbf{f}_{MAL}}{d\ln\mathbf{c}_1} = \frac{\partial\mathbf{f}_{MAL}}{\partial\ln\mathbf{c}_1} + \frac{\partial\mathbf{f}_{MAL}}{\partial\ln\mathbf{c}_2}\frac{\partial\ln\mathbf{c}_2}{\partial\ln\mathbf{c}_1} = 0 \tag{26}$$

Those derivative can be calculated by means of the following linear system

$$\frac{\partial\mathbf{f}_{MAL}}{\partial\ln\mathbf{c}_2}\frac{\partial\ln\mathbf{c}_2}{\partial\ln\mathbf{c}_1} = -\frac{\partial\mathbf{f}_{MAL}}{\partial\ln\mathbf{c}_1} \tag{27}$$

The conversion to $\partial\mathbf{c}_2/\partial\mathbf{c}_1$ is straightforward, recalling that $d\ln x/dx = 1/x$:

$$\frac{\partial c_{2,i}}{\partial c_{1,j}} = \frac{c_{2,i}}{c_{1,j}}\frac{\partial\ln c_{2,i}}{\partial\ln c_{1,j}} \tag{28}$$

The derivatives of (22), remembering the definition of activity ($\mathbf{a} = \boldsymbol{\gamma}\cdot\mathbf{c}$) and that $\boldsymbol{\gamma} = f(\mathbf{c})$, read

$$\begin{aligned}\frac{\partial\varepsilon_i}{\partial\ln c_{1,j}} = \mathrm{L}_{1,ij} &+ \sum_{m=1}^{N_1}\mathrm{L}_{1,im}\frac{\partial\ln\gamma_{1,mj}}{\partial\ln c_{1,j}} + \\ &+ \sum_{l=1}^{N_2}\mathrm{L}_{2,il}\left(\frac{\partial\ln\gamma_{2,lj}}{\partial\ln c_{1,j}} + \frac{\partial\ln c_{2,lj}}{\partial\ln c_{1,j}}\right) \\ i &= 1,\ldots,N_a \\ j &= 1,\ldots,N_1\end{aligned} \tag{29}$$

Matrices $\mathbf{L}_1$ and $\mathbf{L}_2$ are the parts of matrix $\mathbf{L}$ relative to primary and secondary species, respectively.

The derivatives of (20) with respect to the state variables can be calculated from

$$\frac{\partial \varepsilon_i}{\partial \ln c_{1,j}} = \prod_{n=1}^{N_s} {a_n}^{\mathrm{L}_{in}} \cdot$$
$$\cdot \left[\mathrm{L}_{1,ij} + \sum_{m=1}^{N_1} \mathrm{L}_{1,im} \frac{\partial \ln \gamma_{1,mj}}{\partial \ln c_{1,j}} + \sum_{l=1}^{N_2} \mathrm{L}_{2,il} \left(\frac{\partial \ln \gamma_{2,lj}}{\partial \ln c_{1,j}} + \frac{\partial \ln c_{2,lj}}{\partial \ln c_{1,j}} \right) \right]$$
$$i = 1, \ldots, N_a$$
$$j = 1, \ldots, N_1 \tag{30}$$

References

Appelo, C.A.J., Postma, D., 2010. Geochemistry, groundwater and pollution. Taylor & Francis.

Back, W., Hanshaw, B.B., Pyle, T.E., Plummer, L.N., Weidie, A., 1979. Geochemical significance of groundwater discharge and carbonate solution to the formation of caleta xel ha, quintana roo, mexico. Water Resources Research 15, 1521–1535.

Ball, J.W., Nordstrom, D.K., 1991. User's manual for WATEQ4F, with revised thermodynamic data base and test cases for calculating speciation of major, trace, and redox elements in natural waters. US Geological Survey Denver, CO.

Bea, S., Carrera, J., Ayora, C., Batlle, F., Saaltink, M., 2009. Cheproo: A fortran 90 object-oriented module to solve chemical processes in earth science models. Computers & Geosciences 35, 1098–1112.

Bear, J., 1999. Seawater intrusion in coastal aquifers concepts, methods and practices. Springer.

Bethke, C., 2008. Geochemical and biogeochemical reaction modeling. volume 543. Cambridge University Press Cambridge, UK.

Brassard, P., Bodurtha, P., 2000. A feasible set for chemical speciation problems. Computers & Geosciences 26, 277–291.

Cabaniss, S.E., 1997. Propagation of Uncertainty in Aqueous Equilibrium Calculations : Non-Gaussian Output Distributions. Analytical Chemistry 69, 3658–3664.

Cabaniss, S.E., 1999. Uncertainty propagation in geochemical calculations: non-linearity in solubility equilibria. Applied Geochemistry 14, 255–262.

Carrayrou, J., Mose, R., Behra, P., 2002. New Efficient Algorithm for Solving Thermodynamic Chemistry. Environmental and Energy Engineering 48, 894–904.

Charlton, S.R., Macklin, C.L., Parkhurst, D., 1997. Phreeqcia graphical user interface for the geochemical computer program phreeqc. US Geological Survey Water-Resources Investigations Report 9.

Criscenti, L.J., Laniak, G.F., Erikson, R.L., 1996. Propagation of uncertainty through geochemical code calculations. Geochimica et Cosmochimica Acta 60, 3551–3568.

De Simoni, M., Carrera, J., Sanchez-Vila, X., Guadagnini, A., 2005. A

procedure for the solution of multicomponent reactive transport problems. Water resources research 41.

De Simoni, M., Sanchez-Vila, X., Carrera, J., Saaltink, M., 2007. A mixing ratios-based formulation for multicomponent reactive transport. Water Resources Research 43.

Denison, F.H., Garnier-Laplace, J., 2005. The effects of database parameter uncertainty on uranium (vi) equilibrium calculations. Geochimica et cosmochimica acta 69, 2183–2191.

Duan, Z., Li, D., 2008. Coupled phase and aqueous species equilibrium of the H_2O–CO_2–NaCl–$CaCO_3$ system from 0 to 250 c, 1 to 1000bar with nacl concentrations up to saturation of halite. Geochimica et Cosmochimica Acta 72, 5128–5145.

Gupta, H.V., Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resources Research 48.

Gustafsson, J.P., 2011. Visual MINTEQ 3.0 user guide. Royal Institute of Technology: Stockholm, Sweden.

Helgeson, H.C., Kirkham, D.H., 1974. Theoretical prediction of the thermodynamic behavior of aqueous electrolytes at high pressures and temperatures; ii, debye-huckel parameters for activity coefficients and relative partial molal properties. American Journal of Science 274, 1199–1261.

Kräutle, S., 2011. The semismooth newton method for multicomponent reactive transport with minerals. Advances in Water Resources 34, 137–151.

Kräutle, S., Knabner, P., 2005. A new numerical reduction scheme for fully coupled multicomponent transport-reaction problems in porous media. Water resources research 41.

Kulik, D.A., Wagner, T., Dmytrieva, S.V., Kosakowski, G., Hingerl, F.F., Chudnenko, K.V., Berner, U.R., 2013. Gem-selektor geochemical modeling package: revised algorithm and gems3k numerical kernel for coupled simulation codes. Computational Geosciences 17, 1–24.

Mayer, K.U., Frind, E.O., Blowes, D.W., 2002. Multicomponent reactive transport modeling in variably saturated porous media using a generalized formulation for kinetically controlled reactions. Water Resources Research 38, 13–1.

Meeussen, J.C., 2003. Orchestra: An object-oriented framework for implementing chemical equilibrium models. Environmental science & technology 37, 1175–1182.

Merino, E., 1979. Internal consistency of a water analysis and uncertainty of the calculated distribution of aqueous species at 25 c. Geochimica et Cosmochimica Acta 43, 1533–1542.

Molins, S., Carrera, J., Ayora, C., Saaltink, M.W., 2004. A formulation for decoupling components in reactive transport problems. Water Resources Research 40.

Nitzsche, O., Meinrath, G., Merkel, B., 2000. Database uncertainty as a limiting factor in reactive transport prognosis. Journal of contaminant Hydrology 44, 223–237.

Ödegaard-Jensen, A., Ekberg, C., Meinrath, G., 2004. LJUNGSKILE: a program for assessing uncertainties in speciation calculations. Talanta 63, 907–916.

Parkhurst, D.L., Appelo, C., et al., 1999. User's guide to PHREEQC (Version 2): A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. US Geological Survey Denver.

Paz-García, J.M., Johannessona, B., Ottosena, L.M., Ribeirob, A.B., Rodríguez-Marotoc, J.M., 2013. Computing multi-species chemical equilibrium with an algorithm based on the reaction extents. Computers & Chemical Engineering .

Rezaei, M., Sanz, E., Raeisi, E., Ayora, C., Vázquez-Suñé, E., Carrera, J., 2005. Reactive transport modeling of calcite dissolution in the fresh-salt water mixing zone. Journal of Hydrology 311, 282–298.

Saaltink, M.W., Ayora, C., Carrera, J., 1998. A mathematical formulation for reactive transport that eliminates mineral concentrations. Water Resources Research 34, 1649–1656.

Saaltink, M.W., Vilarrasa, V., De Gaspari, F., Silva, O., Carrera, J., Rötting, T.S., 2013. A method for incorporating equilibrium chemical reactions into multiphase flow models for CO_2 storage. Advances in Water Resources 62, 431–441.

Sanz, E., Ayora, C., Carrera, J., 2011. Calcite dissolution by mixing waters:

505 geochemical modeling and flow-through experiments. Geologica Acta 9,
506 67–77.

507 Smith, S.D., Adams, N.W.H., Kramer, J.R., 1999. Resolving uncertainty
508 in chemical speciation determinations. Geochimica et Cosmochimica Acta
509 63, 3337–3347.

510 Steefel, C., Yabusaki, S.B., 2000. OS3D/GIMRT software for modeling
511 multicomponent-multidimensional reactive transport. Technical Report.
512 Pacific Northwest National Lab., Richland, WA (US).

513 Weber, C.L., VanBriesen, J.M., Small, M.S., 2006. A stochastic regression
514 approach to analyzing thermodynamic uncertainty in chemical speciation
515 modeling. Environmental science & technology 40, 3872–3878.

516 Werner, A.D., Bakker, M., Post, V.E., Vandenbohede, A., Lu, C., Ataie-
517 Ashtiani, B., Simmons, C.T., Barry, D.A., 2013. Seawater intrusion pro-
518 cesses, investigation and management: Recent advances and future chal-
519 lenges. Advances in Water Resources 51, 3–26.

520 Wolery, T.J., 1983. EQ3NR: a computer program for geochemical aqueous
521 speciation-solubility calculations. Users guide and documentation. Tech-
522 nical Report. Lawrence Livermore National Lab., CA (United States).

523 Wolery, T.J., 1992. EQ3NR, a Computer Program for Geochemical Aqueous
524 Speciation-solubility Calculations: Theoretical Manual, User's Guide and
525 Related Documentation (Version 7.0). Lawrence Livermore Laboratory,
526 University of California.

Xu, T., Sonnenthal, E., Spycher, N., Pruess, K., 2006. TOUGHREACT - A simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media: Applications to geothermal injectivity and CO_2 geological sequestration. Computers & Geosciences 32, 145–165.

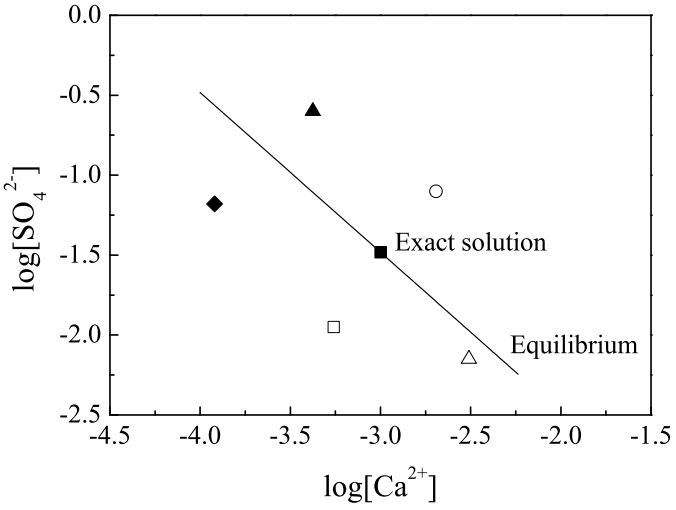

Figure 1: Five measurements, exact solution and equilibrium line for gypsum example.

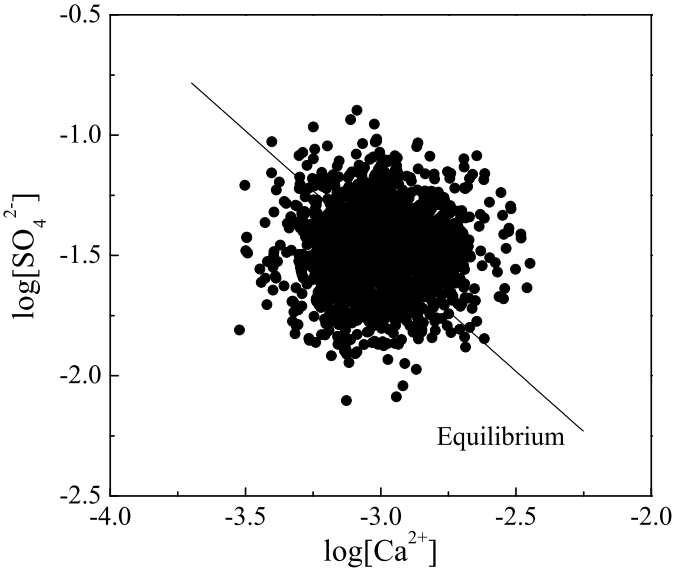

Figure 2: 1500 measurement points generated by means of a lognormal distribution for gypsum example.

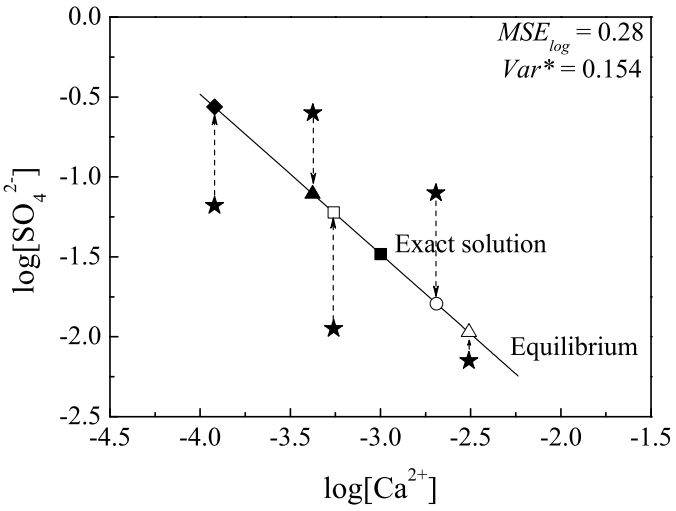

Figure 3: Speciation results of traditional speciation method, exact solution and equilibrium line for gypsum example. Dashed arrows show the movement of the five points from initial conditions, represented with stars.

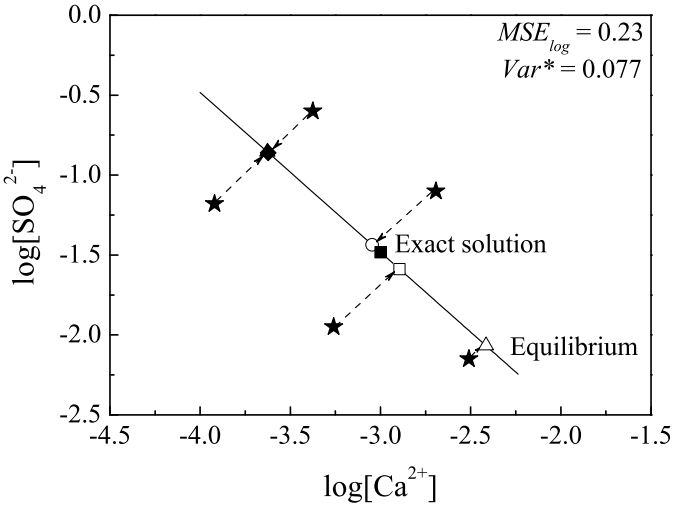

Figure 4: Speciation results of proposed method, exact solution and equilibrium line for gypsum example. Dashed arrows show the movement of the five points from initial conditions, represented with stars.

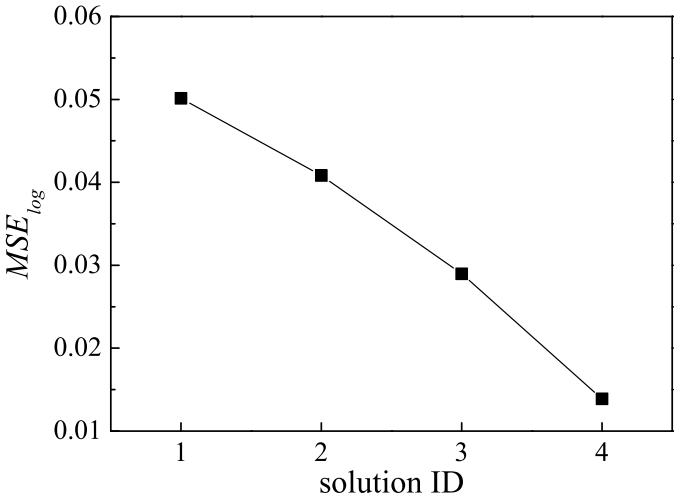

Figure 5: MSE_{log} for traditional speciation method (solution 1) and proposed method (solutions 2, 3 and 4) obtained with σ_e values of table 2.

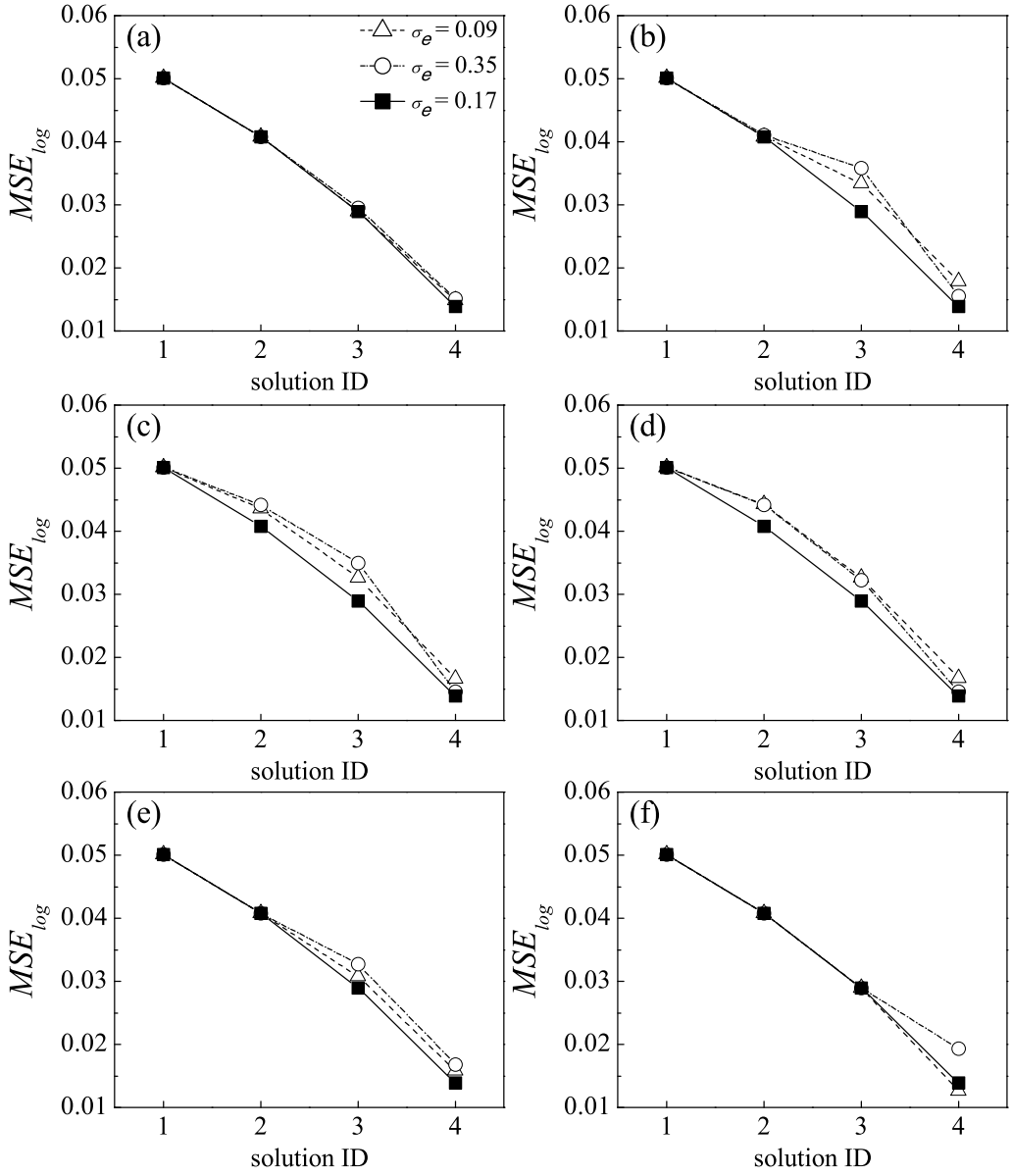

Figure 6: MSE_{log} obtained changing values of σ_e of each data: (a) Ca_{tot}; (b) a_{H^+}; (c) alkalinity; (d) TIC; (e) $a_{CO_{2(g)}}$; (f) calcite equilibrium.

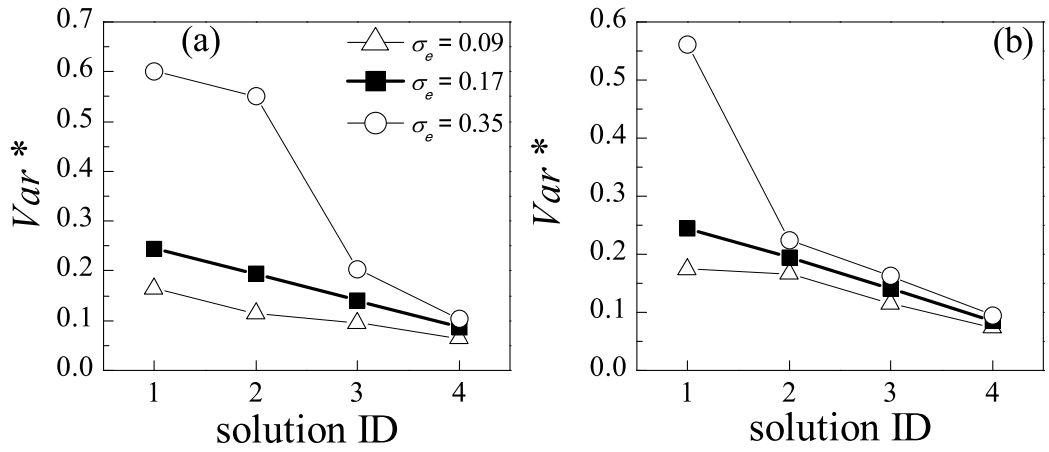

Figure 7: Var^* obtained for different values of σ_e of (a) a_{H^+} and (b) alkalinity.

Table 1: Exact solution for carbonate example.

	Species	Values
c^*	Ca^{2+}	$4.92 \cdot 10^{-4}$
$[mol/l]$	H^+	$5.51 \cdot 10^{-9}$
	HCO_3^-	$9.63 \cdot 10^{-4}$
	$CO_{2(aq)}$	$1.07 \cdot 10^{-5}$
	CO_3^{2-}	$9.78 \cdot 10^{-6}$
	OH^-	$2.01 \cdot 10^{-6}$
a*	$CO_{2(g)}$	$3.16 \cdot 10^{-4}$
$[bar]$		
SI	$Calcite$	0.0
(Sat. Index)		

Table 2: Mean values and standard deviations adopted to generate 1500 realizations of data for the carbonate example.

Data	μ	σ_g	σ_e	ε equation
Ca_{tot}	$4.92 \cdot 10^{-4}$	0.17	0.17	(21)
a_{H^+}	$5.29 \cdot 10^{-9}$	0.17	0.17	(22)
Alkalinity	$9.85 \cdot 10^{-4}$	0.17	0.17	(21)
TIC	$9.84 \cdot 10^{-4}$	0.17	0.17	(21)
$\mathrm{a}_{CO_{2(g)}}$	$3.16 \cdot 10^{-4}$	0.17	0.17	(22)
Calcite Eq.	0	0	0.17	(22)

Units are in mol/l except for $\mathrm{a}_{CO_{2(g)}}$ which is expressed in bar.