Applied Geochemistry 55 (2015) 184–191

Contents lists available at ScienceDirect

Applied Geochemistry

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

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, Netherlands

ARTICLE INFO

Article history:
Available online 20 December 2014

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 or 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 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 results show that using redundant data improves speciation results reducing the estimation error between computations and measurements. In fact, the larger the amount of redundant data, the better the speciation in terms of errors of the estimated concentrations.

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 modeling 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 and Appelo, 1999) and its interactive version, PHREEQCi (Charlton et al., 1997), EQ3NR (Wolery, 1983; Wolery, 1992) and WATEQ4F (Ball and Nordstrom, 1991).

Speciation calculations are subject to 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), Cabaniss (1997), Criscenti et al. (1996) and Merino (1979). Smith et al., 1999

* Corresponding author at: GHS, Department of Geotechnical Engineering and Geosciences, Universitat Politecnica de Catalunya, UPC-BarcelonaTech, c/Jordi Girona 1-3, 08034 Barcelona, Spain.

E-mail 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).

http://dx.doi.org/10.1016/j.apgeochem.2014.12.005
0883-2927/

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 probabilistic methods.

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 acknowledging 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 even further the estimation errors.

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, 1999, 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

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

where [ ] represents molal concentration (mol/kg w). 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 $\mathbf{c}_1$; and (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 acknowledge 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

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