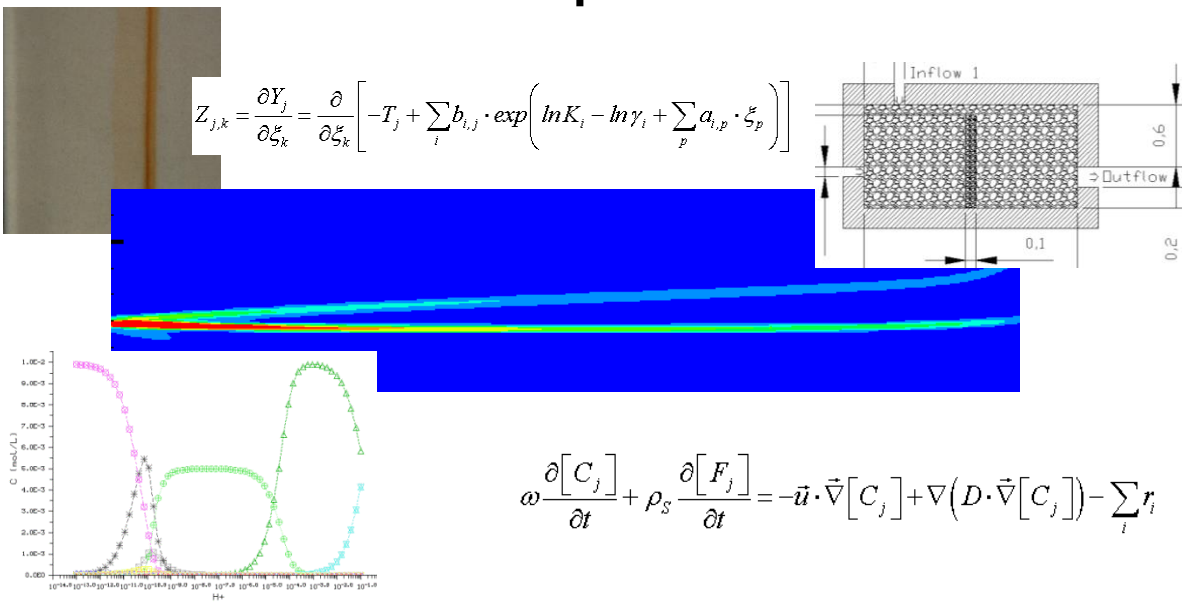
Université de Strasbourg
Laboratoire d'**Hy**drologie et de **Gé**ochimie de **S**trasbourg
UMR 7517 CNRS UdS/EOST ENGEES

Résumé des travaux en vue d'obtenir
l'**Habilitation à Diriger des Recherches**
soutenue le **21 mars 2018** à **Strasbourg**

Modélisation des phénomènes de transport réactif en milieu poreux saturé



$$Z_{j,k} = \frac{\partial Y_j}{\partial \xi_k} = \frac{\partial}{\partial \xi_k}\left[-T_j + \sum_i b_{i,j} \cdot exp\left(lnK_i - ln\gamma_i + \sum_p a_{i,p} \cdot \xi_p\right)\right]$$

$$\omega \frac{\partial [C_j]}{\partial t} + \rho_S \frac{\partial [F_j]}{\partial t} = -\vec{u} \cdot \vec{\nabla}[C_j] + \nabla\left(D \cdot \vec{\nabla}[C_j]\right) - \sum_i r_i$$

Garant : Philippe Acherer, DR. CNRS

Rapporteurs :

Carole Rosier, Pr. Université du Littoral Côte d'Opale

Brahim Amaziane, MCF. HDR. Université de Pau et des pays de l'Adour

Dominique Trébouet, Pr. Université de Strasbourg

Examinateurs :

Jocelyne Erhel, DR. INRIA

Benoit Madé, ANDRA

Michel Kern, CR. INRIA

Jérôme Carrayrou
Maitre de conférence à l'IUT Louis Pasteur de Strasbourg

IUT Louis Pasteur
Institut universitaire de technologie
Université de Strasbourg

Jérôme CARRAYROU
jerome.carrayrou@unistra.fr
Nationalité française, 43 ans

Maître de Conférence à l'IUT Louis Pasteur
Université de Strasbourg

Recherche : **L**aboratoire d'**Hy**drogéologie et de **Gé**ochimie de **S**trasbourg UMR 7517

Cursus universitaire

1995/1998 : Diplôme d'ingénieur de l'ENGEES

1997/1998 : DEA Mécanique et Ingénierie, Université Strasbourg 1.

1998/2001 : Thèse de l'Université Strasbourg 1, direction de MM Ph. Behra et R. Mosé « Modélisation bi- et tridimensionnelle du transport de solutés réactifs en milieu poreux saturé ».

2001/2002 : ATER à l'IUT Louis Pasteur, Université Louis Pasteur Strasbourg 1.

Depuis 2002 : Maître de conférence à l'IUT Louis Pasteur Strasbourg.

Enseignements actuels

Mécanique des fluides (DUT 1ère et 2nde année 153 h ETD)

Mathématiques (DUT 2nde année 35 h ETD)

Pollution de l'eau et de l'air (Master droit de l'environnement 30 h CM)

Activités administratives

Depuis 2006 : Responsable de l'option Génie de l'Environnement au département Génie Biologique de l'IUT Louis Pasteur.

Animation de la recherche

Co-organisation d'un mini-symposium avec M. Kern, INRIA Rocquencourt : Numerical methods for reactive transport : beyond the operator splitting, *SIAM Conference on Mathematical and Computational Issues in the Geosciences, 7-10 juin 2005, Avignon*

Organisation d'un workshop international : International Workshop on Reactive Transport, Strasbourg 21-24 janvier 2008

Développement d'un benchmark de transport réactif pour le GDR MoMaS, collaboration avec l'INRIA et l'ANDRA – resp. J. Carrayrou (2004-2010)

Recherche

Le but de ces recherches est de développer un outil numérique de simulation du transport de solutés réactifs en milieu poreux saturé et de mettre au point une méthodologie d'application d'un tel outil. Ces recherches ont débuté par le développement de l'outil numérique. Aujourd'hui elles sont constituées d'itérations entre développement numérique et exploitation ou mise au point d'expérimentations en laboratoire et sur site.

Publications et communications

15 articles dans des journaux à comité de lecture, Impact factor supérieur à 1
6 articles en premier auteur - 3 articles en dernier auteur avec des étudiants que j'ai encadré.
1 logiciel en cours de dépôt.
11 communications dans des conférences internationales avec actes.
126 citations ; h-index 6 (octobre 2017).

Journaux scientifiques à comité de lecture

Machat H., **Carrayrou J.** (2017). Comparison of linear solvers for equilibrium geochemistry computation. Computational Geosciences. **21**, 131-150. DOI: 10.1007/s10596-016-9600-5

Marinoni M., **Carrayrou J.**, Lucas Y., Ackerer Ph. (2017). Thermodynamic equilibrium solutions through a modified Newton Raphson method. On line AIChE J. DOI: 10.1002/aic.15506

Carrayrou J., Hoffman J., Knabner P., Krautle S., Dedieuleveult C., Erhel J. , Van der Lee J., Lagneau V., Kern M., Amir L, Mayer K.U., McQuarrie K.T.B (2010,c). Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems—the MoMaS benchmark case. Computational Geosciences. **14**, 483-502. DOI: 10.1007/s10596-010-9178-2

Carrayrou J. (2010,b). Looking for some reference solutions for the reactive transport benchmark of MoMaS with SPECY. Computational Geosciences. **14**, 393-403. DOI: 10.1007/s10596-009-9161-y

Carrayrou J., Kern M., Knabner P. (2010,a). Reactive transport benchmark of MoMaS. Computational Geosciences. **14**, 385-392. DOI: 10.1007/s10596-009-9157-7

Majdalani S., Fahs M., **J. Carrayrou J.**, Ackerer Ph. (2009). Reactive transport parameter estimation: Genetic Algorithm versus Monte Carlo approach, *AIChE. J.* **55,** 1959-1968. DOI: 10.1002/aic.11796

Fahs, M., **Carrayrou, J.**, Younes, A., Ackerer, Ph. (2008). On the efficiency of the direct substitution approach for reactive transport problems in porous media. *Water Air Soil Pollution* **193**, 299-308 DOI: 10.1007/s11270-008-9691-2

Belfort B., **Carrayrou J.**, Lehmann F. (2007). An easily implementable adaptive time stepping method: applications to chemistry, reactive transport and unsaturated flow in porous media. *Transport in Porous Media* **69**, 123-138. DOI: 10.1007/s11242-006-9090-3

Aggarwal M., Ndiaye M.C.A, **Carrayrou J.** (*2007*). Parameter estimation for reactive transport: a way to test the validity of a model, *Physics and Chemistry of the Earth, part A/B/C* **32**, 518-529. DOI: 10.1016/j.pce.2005.12.003

Wanko A., Mosé R, **Carrayrou J**., Sadowski A.G. (2006). Simulation of biodegradation in infiltration seepage – Model development and hydrodynamic validation. *Water Air Soil Pollution* **117**, 19-43. DOI: 10.1007/s11270-005-9046-1

Aggarwal M., **Carrayrou J.** (*2006*). Parameter estimation for reactive transport by a Monte-Carlo approach, *AIChE. J.* **52**, 2281-2289. DOI: 10.1002/aic.10813

Carrayrou J., Mosé R., Behra Ph. (*2004*). Efficiency of operator splitting procedures for solving reactive transport equation, *J. Contam. Hydrol.* **68**, 239-268. DOI : 10.1016/S0169-7722(03)00141-4

Carrayrou J., Mosé R., Behra Ph. (*2003*). Modélisation du transport réactif en milieu poreux : schéma itératif associé à une combinaison d'éléments finis discontinus et mixtes-hybrides, *Comptes Rendus Ac. Sci Mécanique* **331**, 211-216. DOI: 10.1016/S1631-0721(03)00040-8

Carrayrou J., R. Mosé, Ph. Behra (*2002*). A new efficient algorithm for solving thermodynamic chemistry. *AIChE. J.* **48**, 894-904. DOI: 10.1002/aic.690480423

Logiciel et brevets

Dépôt de logiciel en cours : Di Chiara Roupert R., Schäfer G., Ackerer P., **Carrayrou J.**, Quintard M., Marcoux M., Côme J.-M., Chastanet J., 2016. cubicM v2.0, code de calcul multiphase, multicomponent & multiprocess. Déclaration de logiciel (déposée le 25/01/2016 au CNRS délégation Alsace) en cours.

Colloques avec actes

Carrayrou J., Loyaux-Lawniczak S., Lehmann F., DiChiara R., Ackerer Ph. (2017) Modeling laboratory scale experiment (TRACE and SPECY) of Fe-Cr redox reaction along with precipitation and porosity change. *MAMERN VII 17-20 mai 2017, Oujda, Maroco*

Machat H., Carrayrou J. (2015) Working with very ill-conditioned matrices during geochemical modelling. *MAMERN VI 1-5 juin 2015, Pau.*

Carrayrou J. (2009) Setup of the MoMaS Benchmark. *SIAM Conference on Mathematical and Computational Issues in the Geosciences, 14-18 juin 2009, Leipzig, Germany.*

Ackerer Ph., Carrayrou J. (2009). On the benchmarking of reactive transport codes: the MoMaS test cases. *TRePro II International Workshop on Modelling of Coupled Transport Reaction Processes. 18 – 19 mars 2009, Karlsruhe, Germany.* Sur invitation.

Carrayrou J. (2008) First comparative analysis of benchmark's results. *International Workshop on reactive transport, 21-24 janvier 2008, Strasbourg.*

Carrayrou J. (2008) SNIA scheme with operator-specific methods applied to the Reactive Transport Benchmark of GdR MoMaS. *International Workshop on reactive transport, 21-24 janvier 2008, Strasbourg.*

Carrayrou J., Lagneau V. (2007) The reactive transport benchmark proposed by GdR MoMaS: presentation and first results, *Eurotherm81, 4-6 juin 2007, Albi.*

Belfort B., Carrayrou J., Lehmann F. (2006) An adaptative time stepping method based on Richardson extrapolation, *IAHR international groundwater symposium on groundwater hydraulics in complex environments, 12-14 juin 2006, Toulouse.*

Carrayrou J. (2005). Overview of operator splitting methods for reactive transport. *SIAM Conference on Mathematical and Computational Issues in the Geosciences, 7-10 juin 2005, Avignon.*

Fahs M., Carrayrou J., Ackerer Ph. (2005). Comparison of two formulation for reactive transport by global approach *SIAM Conference on Mathematical and Computational Issues in the Geosciences, 7-10 juin 2005, Avignon.*

Carrayrou J., Mosé R., Behra Ph. (*2001*). Modélisation des systèmes chimiques à l'équilibre thermodynamique : revue et comparaison, *8ème congrès francophone de génie des procédés, 17-19 octobre 2001, Nancy.*

Mémoires et rapports

Bourgeat, A., Bryant, S., Carrayrou, J., Dimier, A., Duijn, C.V., Kern, M., Knabner, P., Leterrier, N. (2007). GdR MoMaS benchmark reactive transport. http://www.gdrmomas.org/ex_qualifications.html

Impact Factors

Journal	Articles	Impact Factor 2016
AIChE J	4	2,836
Computational Geosciences	4	1,602
Water Air Soil Pollution	2	1,702
Transport in Porous Media	1	2,205
J. Contam. Hydrol	1	2,009
Physics and Chemistry of the Earth	1	1,426
Comptes Rendus Ac. Sci Mécanique	1	1,026

Table des matières

INTRODUCTION

Enjeux sociétaux

La gestion quantitative et qualitative de la ressource en eau est, depuis plusieurs décennies, une préoccupation politique et scientifique majeure. Il est plus que probable que ces préoccupations s'accroissent encore durant le 21ème siècle. En effet, plusieurs facteurs (explosion démographique, modification des habitudes alimentaires, accroissement de l'industrialisation, hausse du niveau de vie et changements climatiques) vont se combiner pour accroitre la tension sur la ressource en eau [1]. Cette tension va nécessiter une gestion plus précise de cette ressource, tant au niveau quantitatif que qualitatif. De plus, des études récentes [2] montrent que les activités humaines sont désormais susceptibles de modifier durablement le sous-sol, qu'il s'agisse de ses propriétés hydrauliques [3], la structure de la matrice solide [4] ou la composition chimique des aquifères [5-7]. Ainsi, la combinaison de ces différents facteurs (besoins croissants, sensibilité plus forte à la qualité des eaux, augmentation des impacts et diversification des polluants) va nécessiter des capacités prédictives nouvelles, portant tant sur la quantité que sur la qualité de la ressource en eau.

L'outil appelé *modèle de transport réactif*, permettrait de prédire l'évolution de la qualité (voire de la quantité) d'une ressource en eau, en intégrant les propriétés géologiques, géochimiques et biologiques du milieu, les différents contaminants et leurs interactions (entre eux et avec le milieu). On comprend aisément qu'un tel outil représenterait un élément prépondérant dans la gestion de la ressource en eau dans un contexte de pression croissante.

Les récents développements des modèles de transport réactif dédiés à l'étude des hydrosystèmes naturels permettent aujourd'hui d'envisager une gestion prévisionnelle de la ressource en eau [2, 8, 9]. Les domaines d'application des modèles de transport réactif sont aujourd'hui extrêmement vastes. Historiquement, les premiers travaux ont étudié l'impact du transfert sur les modifications de composition des eaux [10], puis sur celui de contaminants minéraux tels que des métaux lourds [11]. La problématique du devenir des déchets nucléaires à longue durée de vie et l'étude du stockage souterrain de ceux-ci a ouvert un très vaste champ de tests et de développement [12-14]. Les enjeux liés au réchauffement climatique ont posé la question de la faisabilité du stockage de CO_2 ; ceci a donné lieu à de nombreux travaux [15-19]. De nombreuses études s'intéressent désormais au devenir des nouveaux contaminants dans l'environnement, parfois avant l'apparition d'éventuelles crises sanitaires : nanoparticules [20-22], nouvelles molécules phytosanitaires [23-25], médicaments [26, 27], solvants organiques [28]… On voit apparaitre de plus en plus de travaux portant sur la modélisation de systèmes aquifères complexes. La complexité peut alors résider dans l'hydrodynamique [29], dans la structure géologique ou géométrique de l'aquifère [30], dans la géochimie ou la biogéochimie impliquée [16, 17, 31-36], ou encore dans la diversité des phénomènes pris en compte : hydrodynamique, thermique, mécanique, chimique [32, 37].

Contexte scientifique

Lorsque l'on parle de modèle de transport réactif en géosciences, le choix des phénomènes décrits est extrêmement vaste. Les premiers modèles présentaient des domaines monodimensionnels [10, 38, 39]. Actuellement, les domaines décrits sont bi- ou tridimensionnels [40-45] avec des milieux hétérogènes. Certains modèles prennent en compte la présence de fractures [41, 46-48], d'autres incluent des écoulements multiphasiques [49-52]. Concernant la partie réactivité, les modèles ont débuté avec une chimie en phase aqueuse et/ou des phénomènes de sorption par isotherme [53, 54]. Puis des phénomènes chimiques plus complexes ont été pris en compte : échange d'ions [55, 56], précipitation-dissolution [40, 57-59], redox [5, 28, 60], complexation de surface [61-63], formation de solutions solides [64], prise en compte de solutions non idéales [65]. Les phénomènes biologiques sont aujourd'hui inclus dans de nombreux travaux, soit par le biais de microorganismes [66-68] ou de phénomènes racinaires [69, 70]. Enfin, de nombreuses rétroactions sont prises en compte, décrivant les conséquences des phénomènes chimiques sur la structure du milieu poreux ou sur les caractéristiques du fluide. Certains travaux décrivent l'évolution des propriétés hydrodynamiques du milieu poreux : porosité, tortuosité, dispersivité [40, 71-74]. D'autres s'intéressent à la production de chaleur due aux réactions (souvent décroissance radioactive) et aux conséquences de l'élévation de température [32, 37, 40]. Enfin, il est possible d'intégrer les contraintes mécaniques s'exerçant sur le milieu et de décrire les déformations de celui-ci ainsi que l'évolution de ses propriétés mécaniques [32, 75, 76].

Un modèle de transport réactif, pour être utilisable, doit être construit de plusieurs éléments : un modèle, une formulation mathématique de ce modèle, des méthodes numériques, un code informatique dans un langage donné, des paramètres nécessaires au modèle et une description de l'état initial et des conditions aux limites du domaine étudié. On comprend alors que cet outil, essentiel pour l'analyse du comportement des hydro-systèmes [77] doit obéir à certaines règles, et présente des limitations intrinsèques [78].

(i) Ces codes de transport réactif sont basés sur des modèles conceptualisant les phénomènes de transport réactif dans les hydrosystèmes. En conséquence de quoi, ces codes ne peuvent être considérés comme *vrais*. Un modèle n'est jamais qu'une image idéalisée d'un système et, s'il est possible de prouver qu'un modèle est faux, la réciproque est fausse. Bredehoeft et Konikow [79, 80] soulignent que le terme *validation*, très largement usité, devrait être exclu car ce terme a, dans l'esprit du public, une connotation de vérité très forte.

(ii) On distingue habituellement modèles heuristiques et mécanistes. Un modèle heuristique est sensé, après calage, simuler le comportement d'un système dans des conditions similaires à celles de son calage. Si les conditions ne sont pas conformes, un tel modèle n'est pas applicable. De plus en plus, nous développons des modèles mécanistiques qui devraient être capable de s'affranchir de cette limite. Cependant, un modèle mécanistique dépend de paramètres indispensables à la description du système modélisé. Or pour des systèmes complexes, l'acquisition de ces paramètres dépend très souvent d'un processus de calage. Ainsi que l'a montré Brusseau [81], il est tout à fait possible de caler de façon satisfaisante un modèle inadapté. A ce niveau de complexité,

la frontière entre modèle heuristique et mécanistique s'estompe et un tel modèle mécanistique dépend alors fortement des conditions ayant présidées à son calage.

Les challenges qui se présentent aujourd'hui pour la partie conceptuelle de la modélisation (les modèles et leurs formulations mathématiques) vont porter sur les éléments suivants :

(i) L'intégration de la bio-géochimie [9]. Il s'agira non pas d'associer des modèles semi-empiriques de type loi de Monod à des modèles géochimiques, mais d'acquérir une compréhension des phénomènes biologiques permettant de développer des modèles mécanistiques, suffisamment souples et universels pour rester valides dans des conditions hydrodynamiques et chimiques variables.

(ii) Augmenter la complexité de la partie hydrologique [9] pour intégrer les phénomènes de changement environnemental ou la pression sociale.

(iii) Intégrer les modifications irréversibles observées [2], conséquence des actions anthropiques sur la structure physico-chimique des hydro systèmes.

(iv) Gérer les différentes échelles [77], afin de surmonter les écarts reportés entre expériences de laboratoires et valeurs au champ, et pouvoir passer d'une réactivité décrite à l'échelle du pore à une échelle au moins décamétrique.

(v) Assurer le couplage entre les processus chimiques et mécaniques en sub-surface [77]. Ce point nécessitera de gérer efficacement les différentes interactions et rétroactions entre les différentes modifications de la phase fluide, de la matrice solide et des propriétés des réactions chimiques [3, 76].

Concernant les méthodes numériques et la partie informatique, les verrous actuels dépendent fortement de l'évolution du matériel informatique [82]. Au cours des 20 dernières années, les conditions matérielles qui ont présidé à l'essor de cette discipline [83], ont évolué avec des coûts plus abordables, des ressources mémoires plus importantes et des capacités de calculs plus rapides. Cependant, si les coûts vont certainement continuer à baisser et les capacités de mémoire à augmenter, il semble que la vitesse de calcul soit arrivée à un plafond.

(i) A l'heure actuelle, les progrès en ce domaine portent sur la multiplication de processeurs ou de cœurs. Ainsi les futurs développements des codes devront s'adapter à cette nouvelle tendance et les efforts devront porter sur la parallélisation des codes [82].

(ii) La baisse des coûts de la mémoire et l'augmentation de la capacité de mémoire disponible a supprimé les contraintes relatives aux nombre de variables, permettant la prise en compte de systèmes extrêmement complexes. Cette complexité peut s'exprimer par la description de domaines vastes et hétérogènes ou/et par la prise en compte d'une grande variété de phénomènes. L'augmentation des capacités mémoire a eu également pour conséquence de rendre caduc l'un des principaux défauts des méthodes de résolution globales : leur grande consommation en place mémoire [83]. On a montré [84] qu'aujourd'hui, les méthodes globales sont au moins aussi performantes que celles par séparation d'opérateurs.

(iii) Cependant, si les approches globales ont aujourd'hui tendance à être aussi performantes, sinon davantage, que les approches par séparation d'opérateurs, elles présentent encore un inconvénient de taille. Dans le cas d'une approche globale, il est en

effet nécessaire d'intégrer en un seul jeu d'équations l'intégralité des phénomènes pris en compte. Ceci nécessite un travail de codage extrêmement lourd, tant lors du développement initial que lors des évolutions du code. Les approches par séparation d'opérateurs, en permettant un développement distinct de plusieurs éléments de code sont plus simples à écrire, coder et maintenir [85].

Travaux réalisés

Développement du code

Ces recherches portant sur la modélisation du transport réactif ont débuté à l'Institut de Mécanique des Fluides et des Solides (UMR 7507 ULP-CNRS) de Strasbourg [86]. S'appuyant sur les compétences acquises au sein de l'unité, un modèle par séparation d'opérateurs a été développé sur la base d'une combinaison d'éléments finis discontinus et d'éléments finis mixtes hybrides pour la résolution de l'opérateur de transport [87]. Cette combinaison s'est révélée très efficace, permettant de minimiser les phénomènes de diffusion numérique induit par l'approche par séparation d'opérateurs, tout en fournissant une solution numérique stable.

Il est cependant reconnu que l'approche par séparation d'opérateurs induit des erreurs numériques, liées à la méthode de résolution [83, 88]. Une étude analytique basée sur le suivi des bilans de masse nous a permis de quantifier ces erreurs de séparation d'opérateurs dans le cas d'une chimie décrite par loi cinétique [89]. On a montré que ces erreurs dépendent de la vitesse de réaction et du pas de temps choisi. En conduisant cette étude sur les différents schémas de séparation d'opérateurs utilisés, nous avons pu sélectionner les schémas les plus performants. Afin de vérifier numériquement ces résultats, et pour disposer d'un outil de référence, un code de transport réactif par approche globale a été développé lors de la thèse de M. Fahs. Au cours de ce travail, une étude systématique des diverses approches globales possibles a été menée pour montrer que l'approche par substitution est la plus efficace [90].

L'approche par séparation d'opérateurs qui a été choisie nécessite la résolution de l'opérateur de chimie, écrit dans le cas de l'équilibre instantané comme un système algébrique non linéaire. Ce système résulte de la combinaison des lois d'action de masse et de conservation [91]. Classiquement, de tels systèmes sont résolus par des méthodes de type Newton-Raphson. Néanmoins, ces méthodes ne sont pas assez robustes pour des systèmes aussi fortement non linéaires que ceux décrivant les phénomènes chimiques. L'étude de ces problèmes de robustesse a permis de développer de nouvelles méthodes de résolutions, comme la méthode des fractions continues positives ou la définition d'un domaine définition chimique [92]. L'implémentation de ces nouveaux outils nous a permis d'obtenir un code particulièrement robuste et rapide comme cela a été prouvé par la suite [84].

Dans le cas où l'opérateur de chimie est décrit par lois cinétiques, les temps caractéristiques des différents phénomènes, transport et chimie, sont très différents. De plus, il est fréquent que les temps caractéristiques des cinétiques chimiques changent au cours de l'évolution d'un même système. Afin de capturer ces évolutions, il est nécessaire d'avoir un pas de temps de calcul suffisamment petit pour suivre le plus rapide des phénomènes. Mais, dans le cas d'une résolution par séparation d'opérateurs, il n'est pas forcément nécessaire d'utiliser le plus petit pas de temps pour tous les opérateurs de calcul. Nous avons montré que la méthode d'extrapolation de Richardson

permet une adaptation du pas de temps basée sur un estimateur d'erreur à postériori efficace [93]. Cette méthode s'est révélée transposable à diverses thématiques : cinétique chimique, écoulement en milieu non saturé, transport réactif.

Applications

Les outils et méthodes développés ont conduit à un code flexible, robuste et rapide : SPECY. Ces qualités ont été mises à profit pour modéliser divers systèmes hydro-géo-chimiques. Ainsi ce code a été adapté à la description de processus d'épuration des eaux par infiltration-percolation [66]. Les phénomènes pris en compte pour décrire ce système sont complexes : écoulement transitoire non saturé, transport de l'oxygène gazeux, transport des nutriments dissous, cinétique de croissance du biofilm, de consommation de l'oxygène et de dégradation des nutriments. La capacité du code à intégrer aisément ces différents éléments montre la validité des choix qui ont été faits.

Estimation de paramètres

La rapidité et la robustesse du code nous a également permis d'initier une nouvelle approche : l'estimation de paramètres en transport réactif. Classiquement, les paramètres chimiques sont estimés à partir d'expériences faites en réacteur fermé [94]. Or ce type d'expérience ne permet pas toujours d'appréhender l'intégralité des phénomènes chimiques mis en jeu. A cause de la forte non-linéarité des phénomènes étudiés, nous avons choisi de développer une méthode d'estimation de paramètres en transport réactif basée sur une approche de type Monte-Carlo [62]. Les paramètres estimés sont les paramètres spécifiques aux phénomènes chimiques : constantes d'équilibres, concentrations en sites de fixation… L'estimation de paramètres est réalisée sur un jeu de plusieurs expériences [95], afin d'obtenir des valeurs applicables à un vaste ensemble de conditions expérimentales. En testant différentes configurations réactionnelles, nous avons montré [63] que l'approche par estimation de paramètres permet de sélectionner le ou les mécanismes réactionnels les plus probables, et que la sélection obtenue est en accord avec les éléments théoriques connus. Cependant, les temps de calculs nécessaires à ce type d'approche sont très longs. La méthode initialement choisie, de type Monte Carlo, est très gourmande en temps de calcul. Une autre approche a été tentée, une approche par algorithme génétique [96], qui s'est révélée plus rapide que l'approche Monte-Carlo.

Comparaison de méthodes

L'expérience acquise concernant les enjeux numériques liés à la modélisation du transport réactif a été mise à profit pour développer, dans le cadre du GdR MoMaS, une série d'exercices dédiés à la comparaison des codes de transport réactif. La conception de ces exercices a répondu à un cahier des charges assez délicat : présenter un niveau de difficulté suffisant pour être discriminant vis-à-vis des outils numériques utilisés, mais être suffisamment simple d'un point de vue hydrodynamique et chimique pour être accessible à une communauté assez vaste [97]. Le choix s'est donc porté sur une chimie idéale, avec un petit nombre d'espèces et de réactions possibles, mais avec une très forte non linéarité. L'hydrodynamique est proposée en 1D ou 2D, et le domaine est hétérogène tant d'un point

de vue hydrodynamique que chimique. Enfin, 3 niveaux de difficultés sont proposés : un niveau *easy* pour lequel les phénomènes chimiques sont à l'équilibre instantané sans précipitation-dissolution ; un niveau *medium* avec des phénomènes chimiques décrits par équilibre instantané et d'autres par cinétique, toujours sans précipitation dissolution ; et un niveau *hard* avec des phénomènes chimiques avec précipitation-dissolution décrits par équilibre instantané et d'autres par cinétique [98]. Le niveau de difficulté requis pour obtenir un test discriminant exclu immédiatement la possibilité d'obtenir une solution analytique au problème. Il a donc été nécessaire de rechercher numériquement des solutions de référence afin de pouvoir procéder à l'analyse des résultats. Ces solutions de référence ont été obtenues à l'aide du code SPECY, en procédant à des raffinements de maillage successifs associés à des réductions de pas de temps [99]. La série d'exercices a été soumise à la communauté scientifique et un workshop international a été organisé (*Internationnal workshop on reactive transport, Strasbourg 21-24 janvier 2008*). Ce workshop a été suivi d'un intense travail d'analyse et de synthèse [100]. Les résultats de 7 codes différents ont été étudiés sur la base de la qualité des résultats fournis et des temps de calculs nécessaires [84]. Une très vaste palette de méthodes numériques est représentée dans ce travail :

(i) Approche par séparation d'opérateurs non itérative [99] est utilisée par le code SPECY.
(ii) Approche par séparation d'opérateurs itérative [101] est utilisée par le code HYTEC
(iii) Une approche globale, basée sur une méthode Newton-Krylov est proposée par Amir et Kern [102], qui permet de garder deux entités de code distinctes pour le transport et la chimie.
(iv) Approche globale basée sur une méthode of lines (MOL) est proposée par de Dieuleveult et Erhel [103].
(v) Une approche globale basée sur une méthode par discrétisation-substitution est proposée par le code MIN3P [104].
(vi) Une approche globale, développée pour un code parallèle, et basée un schéma de réduction des variables couplées est proposée par Hoffman et *al.* [105].

Ce travail a montré la robustesse et la qualité des codes développés depuis quelques années (HYTEC, MIN3P, SPECY) ainsi que la validité des nouvelles approches développées. De plus, la supériorité des approches par séparation d'opérateurs, validée dans les années 1990 par les travaux de Yeh et Tripathi [83], se trouve aujourd'hui invalidée. Ce point, justifié par l'évolution des matériels (augmentation de la place mémoire) et des méthodes, est confirmé par ce travail de comparaison.

Objectifs de recherche

Au cours des travaux effectués, j'ai abordé les problèmes de séparation d'opérateurs, de modélisation des phénomènes chimiques, de modélisation numérique d'expériences de laboratoire. L'enjeu général est d'étudier les problématiques multi-phénomènes de transport réactif en milieu poreux afin de comprendre, expliquer et prédire l'évolution d'un milieu poreux ou des contaminants impliqués.

Les premiers travaux impliquaient des phénomènes de transport réactif sans modification structurelle du milieu poreux (porosité constante). Les développements récents mettent en jeu des phénomènes de colmatage impliquant des modifications de porosité, perméabilité et dispersivité du milieu poreux, ainsi que la prise en compte du fractionnement isotopique lors de certaines réactions

(thèse de Marianna Marinoni). Il devient nécessaire d'inclure la possibilité d'affecter des coefficients de diffusion moléculaires spécifique à chaque espèce tout en garantissant le respect de la condition d'électro-neutralité. A plus long terme, l'objectif est d'inclure les transferts de chaleur dans le modèle.

Présentation de ce mémoire

Ce mémoire est composé de 4 parties, la première consacrée à la présentation des phénomènes physico-chimiques et des modèles mathématiques impliqués dans le transport réactif en milieu poreux ; la seconde aux méthodes numériques utilisées et développées pour résoudre ces modèles mathématiques. La troisième partie présente l'exercice de comparaison des codes de transport réactif développé pour le GdR MoMaS et la quatrième est consacrée à deux axes de recherches que je souhaite approfondir et développer.

L'objectif de la première partie de ce mémoire est de faire une présentation assez exhaustive des phénomènes et modèles mathématiques impliqués dans la problématique du transport réactif en milieu poreux. Cette présentation ne sera pas restreinte aux éléments déjà abordés au cours de mes recherches, mais intègrera les éléments en projet, qu'il s'agisse de projets à cours ou à long terme. Néanmoins, nous nous restreindrons aux phénomènes suivants : domaines mono- bi- ou tri-dimensionnel, hétérogène non fracturé, saturé en eau. On supposera le milieu poreux indéformable. Aucune restriction ne sera faite sur les phénomènes chimiques possibles, et même si cette partie n'a pas encore été abordée dans le code développé, nous intègrerons les phénomènes de production et de transfert de chaleur à cette présentation. Un effort particulier sera fait pour présenter les différentes rétroactions et interactions entre les différents phénomènes et les évolutions des propriétés du milieu poreux ou du fluide.

La méthodologie de résolution sera abordée en seconde partie, en présentant les choix techniques et leurs conséquences. Tout d'abord, ce travail s'appuie sur les avancées déjà établies au laboratoire, entre autre sur la modélisation de l'écoulement et du transport en milieu poreux par éléments finis discontinus et mixtes-hybrides [106-108]. Nous montrons [87] que ces méthodes sont également performantes dans le cas du transport réactif multi-espèces. Nous avons choisi de construire le code de transport réactif sur une méthodologie de séparation d'opérateurs. Les méthodes de séparation d'opérateurs ainsi que les erreurs liées à cette séparation d'opérateurs ont été étudiées de façon approfondie [89] et seront présentées ensuite. Enfin, nous aborderons la problématique de la résolution numérique des systèmes chimiques en présentant tout d'abord la spécificité des systèmes à l'équilibre instantané en détaillant la méthode classique de Newton-Raphson, ses limites et les améliorations apportées. Ensuite, nous montrerons comment associer description cinétique et à l'équilibre des phénomènes chimiques avec un algorithme basé sur l'extrapolation de Richardson [93].

En troisième partie seront présentés le contexte, la méthodologie de développement et de valorisation de l'exercice de comparaison des codes de transport réactif développé pour le GdR MoMaS.

En quatrième partie, je présenterai deux points que je souhaite approfondir dans les prochaines années. Le premier, l'estimation de paramètre en transport réactif, a déjà été abordé lors des stages

de M2 de M. Aggarwal et de M. Ndiaye. Il s'agit d'utiliser des expériences de transport réactif pour déterminer les paramètres des lois chimique (constantes d'équilibres, nombre de site de sorption..). Le second, la problématique du maintien de l'électro-neutralité, notamment dans le cas des phénomènes diffusifs, est important pour renforcer les interactions au sein du laboratoire. Le travail entamé avec la thèse de M. Marinoni pour inclure le fractionnement isotopique dans le code donne de bons résultats et les premiers articles sont en préparation. Cependant, de nombreux travaux du laboratoire portent sur des problématiques d'altération où les phénomènes de transport sont fortement diffusifs. Dans ces cas là, les différents coefficients de diffusion induisent un non respect de l'électro-neutralité qu'il est indispensable de controler pour décrire correctement l'évolution du système.

1. PRESENTATION DES PHENOMENES

Les phénomènes décrits dans les modèles de transport réactifs actuels sont connus et décrits, pour la plupart, depuis une vingtaine d'années [109]. Les premiers travaux portent sur la description de systèmes simplifiés [10, 38, 39, 57, 110-112]. Les milieux décrits sont souvent homogènes ; la chimie est idéalisée et souvent un seul phénomène d'immobilisation est étudié (échange d'ions le plus souvent). Néanmoins, une formulation extrêmement complète est rapidement proposée par P. Lichtner [49, 109, 113, 114]. Les éléments sont là pour décrire un système comportant des phénomènes chimiques en phase aqueuse (acide-bases, complexation, oxydation-réduction) décrits à l'équilibre ou sous forme cinétique. Les phénomènes d'immobilisation sont décrits par échange d'ions, sorption ou précipitation-dissolution sous forme d'équilibre ou cinétique. Le transport y est décrit par une équation de Nernst-Plank étendue, comme la superposition d'advection, diffusion (spécifique à chaque ion), dispersion, gradient d'activité, électrocinétique. Les variations de composition minéralogique du milieu vont induire des modifications de ses propriétés hydrodynamiques. Les apports récents portent essentiellement sur le couplage des phénomènes entre eux. Certains phénomènes ont été ajoutés : complexation de surface [51, 115], phénomènes biologiques [116, 117], écoulement en milieu non saturé [37, 50, 51] et d'autres le seront encore par la suite. Mais la principale difficulté réside dans la description des interactions entre tous ces phénomènes.

Nous présenterons ici une formulation basée sur un modèle continu du milieu poreux. Un tel modèle est défini par Lichtner [114] : *Le milieu poreux est représenté comme un continuum dans lequel les phases, gazeuse, solides et fluide coexistent simultanément en chaque point d'un espace mathématique idéal. Les variables spatiales représentant des quantités comme la température, la pression, la concentration en soluté ou la quantité en minéraux sont moyennées sur un volume élémentaire représentatif.* L'objectif de ce chapitre est de donner une vision aussi claire que possible des différents phénomènes impliqués et de leurs interactions. Une synthèse visuelle de ce travail est donnée en Figure 1-1. Cette carte est centrée sur les deux entités en interactions : le milieu poreux et le fluide. Ces interactions vont générer trois types de phénomènes : l'écoulement de la phase fluide, le transport des solutés et de la chaleur ainsi que des réactions chimiques. On peut trouver une représentation de la problématique du transport réactif en milieu poreux utilisant le même outil de communication dans l'article de synthèse de Steefel *et al.* [77]. La présentation choisie par ces auteurs, plus formelle, met en évidence les différents types d'équations : équations de conservation, de l'énergie, des moments, des masses fluides, solides et solutés ainsi que les équations d'état. En centrant la présentation sur les entités physiques plutôt que sur la nature des équations, on met en valeur les interactions entre les phénomènes.

Nous présenterons en premier lieu le milieu poreux et les liens entre les différentes grandeurs qui le décrivent. L'évolution de la structure physique et de la composition minéralogique du milieu, en influant sur l'écoulement de la phase fluide ainsi que sur les différentes réactions chimiques, est l'un des éléments clés d'un modèle de transport réactif. C'est également l'un des verrous scientifiques majeurs.

Ensuite une présentation similaire de la phase fluide sera faite. Issus des domaines scientifiques *classiques* tels que la chimie des solutions et la mécanique des fluides, les influences réciproques entre les différentes caractéristiques du fluide sont plutôt bien connues. On mettra en évidence les liens entre les propriétés du milieu poreux et celles de la phase fluide.

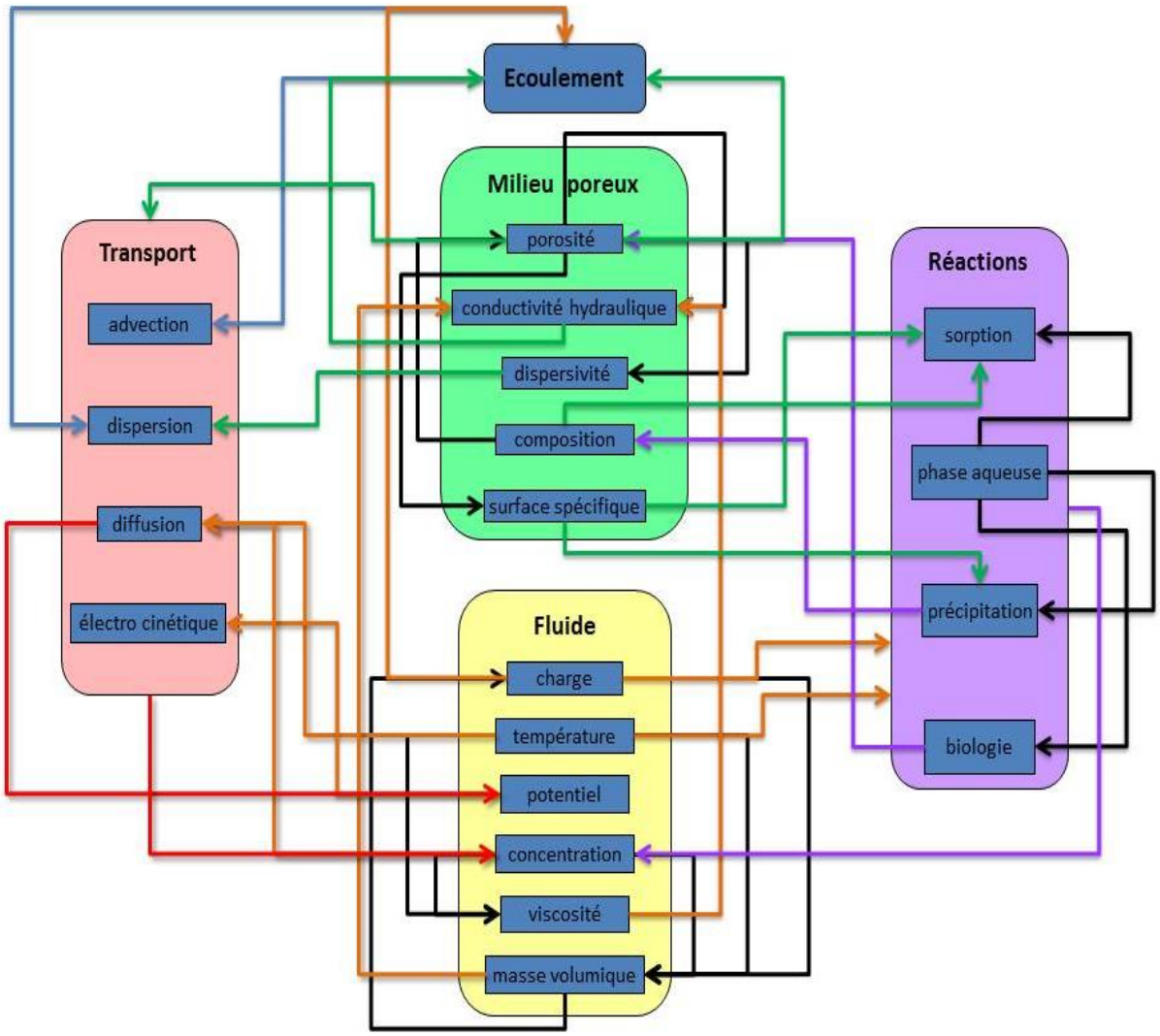


Figure 1-1 : Carte synthétique des interactions lors du transport de soluté réactif en milieu poreux saturé.

Les équations de conservation ; conservation de l'énergie, de la quantité de mouvement, de la quantité de matière ; permettent de décrire les phénomènes de transfert en milieu poreux. Ces relations seront présentées sous leur forme adaptée à la description macroscopique d'un milieu poreux. L'accent sera mis sur les inter-relations entre ces équations et entre les différentes propriétés du milieu poreux et de la phase fluide.

Enfin nous présenterons les principaux phénomènes chimiques rencontrés lors des interactions eau-roches en milieu poreux. Cette présentation n'inclura pas les spécificités propres à chaque minéral, changeantes selon la forme cristallographique adoptée, ni une description étendue des différentes cinétiques chimiques rencontrées. L'objectif est de rester assez synthétique pour focaliser la présentation sur le cœur de ce travail : la modélisation.

1.1. LE MILIEU POREUX

Un milieu poreux est représenté par différents champs de paramètres pouvant évoluer dans l'espace (milieu hétérogène) et/ou dans le temps (milieu évoluant). Afin de décrire un milieu poreux, il faut en premier lieu définir la taille d'un échantillon. Ce Volume Elémentaire Représentatif (VER) est largement défini et discuté dans les ouvrages de référence [118, 119]. Ces notions liées à la taille de l'élément observé sont un point parfois critique en transport réactif, notamment lorsque l'on superpose des phénomènes électrostatiques (complexation de surface, apparition de potentiels de surface) et des phénomènes de transport. En effet, les phénomènes de transport (dispersion entre autre) sont généralement décris à l'échelle d'un VER, souvent d'une taille d'au moins une dizaine de grains. Au contraire, les phénomènes de complexation de surface sont décrits en faisant intervenir la formation d'un potentiel électrostatique à la surface d'un grain. Cette contradiction sera discutée par la suite, et en attendant, nous poursuivrons notre présentation en considérant valide la description par VER. Nous supposerons également que la température (T) et la pression (p) est uniforme à l'échelle du VER, identique pour toutes les phases.

1.1.1. DEFINITION DES GRANDEURS

1.1.1.1. Porosité : ω (-)

La porosité ω d'un milieu poreux est donnée par la fraction du volume vide par rapport au volume total :

$$\omega = \frac{V_{vide}}{V_{total}} \tag{1.1}$$

Cette définition très générale peut être modulée suivant la géométrie du volume vide du milieu. On peut ainsi être amené à utiliser la porosité connectée plutôt que la porosité totale lorsque certains vides du milieu ne sont pas accessible à l'écoulement (voir [118] ou [119] pour une présentation détaillée).

1.1.1.2. Tortuosité : τ (-)

La tortuosité τ d'un milieu poreux est définie comme le ratio entre la longueur du trajet réel Δl d'une particule fluide et la plus courte distance Δx entre deux points (voir Figure 1-2) :

$$\tau = \frac{\Delta l}{\Delta x} \tag{1.2}$$

On peut trouver d'autres formulations pour des concepts similaires, ou d'autres définitions de la tortuosité. Elle est parfois définie comme le carré du rapport Δl sur Δx. On utilise [120, 121] un facteur de résistivité ou facteur de formation, qui a les mêmes caractéristiques. Shen et Chen [121] proposent une revue assez compète des différentes formulations utilisées. Lorsque le milieu est

anisotrope, la tortuosité est alors définie sous forme tensorielle avec une valeur différente selon les axes principaux du milieu [118].

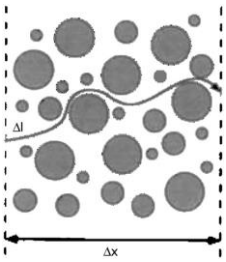


Figure 1-2 : Schéma d'un milieu poreux illustrant le concept de tortuosité
(source : Shen et Chen 2007 [121])

D'après Boudreau [122], la valeur de la tortuosité répond à certaines contraintes géométriques :

(i) La distance moyenne parcourue par un soluté en présence d'un milieu poreux est au moins aussi grande que celle parcourue en eau libre. Cela se traduit par :

$$\tau \geq 1 \tag{1.3}$$

(ii) Lorsque la porosité augmente, la distance parcourue en présence de milieu poreux tend à égaler celle parcourue en eau libre :

$$\lim_{\omega \to 1} \tau = 1 \tag{1.4}$$

(iii) Lorsque le milieu poreux tend à être totalement colmaté, ou imperméable, la distance nécessaire pour faire traverser ce milieu à un soluté devient infinie :

$$\lim_{\omega \to 0} \tau = +\infty \tag{1.5}$$

1.1.1.3. Conductivité hydraulique $\mathbf{K}$ ($m.s^{-1}$) - Perméabilité $\mathbf{k}$ (m^2)

La conductivité hydraulique $\mathbf{K}$ d'un milieu poreux est introduite par la loi empirique proposée par H. Darcy en 1856 [123]. Historiquement présentée sous sa forme monodirectionnelle, elle est aujourd'hui utilisée sous forme généralisée aux problèmes bi- ou tridimensionnels [118] :

$$\omega \vec{u} = -\mathbf{K} \cdot grad(H) \tag{1.6}$$

où $\vec{u}$ ($m.s^{-1}$) est la vitesse du fluide et H (m) la charge hydraulique. La conductivité hydraulique n'est pas une propriété propre au milieu poreux car elle dépend également du fluide. La perméabilité, $\mathbf{k}$, ou perméabilité intrinsèque est, elle, propre au milieu poreux. Elle est liée à la conductivité hydraulique par la relation suivante :

$$\mathbf{K} = \mathbf{k} \frac{\rho_f g}{\mu_f} \tag{1.7}$$

où ρ_f (kg.m^{-3}) est la masse volumique du fluide, μ_f (kg.m^{-1}.s^{-1}) sa viscosité dynamique et g (m.s^{-2}) l'accélération de la pesanteur.

La perméabilité n'est pas une propriété locale du milieu poreux. Elle est la conséquence d'une description à grande échelle (échelle d'un VER) de processus agissant à une échelle beaucoup plus petite (échelle du pore).

1.1.1.4. Dispersion $\overline{\overline{D_i}}$ (m^2.s^{-1}), dispersivité α (m) et diffusion moléculaire d_i (m^2.s^{-1})

Lors du transport d'un soluté *i* en milieu poreux, on constate un étalement du profil de concentration autour de la valeur moyenne. Cet étalement est dû à la superposition de la diffusion moléculaire dans le milieu poreux d_i et de la dispersion mécanique D_m. La diffusion moléculaire est liée à l'agitation thermique des molécules du solvant et du soluté *i*. La dispersion mécanique est causée par l'hétérogénéité locale du champ de vitesse. On écrit le tenseur de dispersion comme la somme de ces deux termes [40] :

$$\overline{\overline{D_i}} = D_m + d_i \tag{1.8}$$

Etant liée au champ de vitesse, la dispersion mécanique change avec celui-ci. Bear [118] donne la relation suivante :

$$\overline{\overline{D_i}} = \left(\alpha_T \omega |u| + d_i\right) \cdot Id + \left(\alpha_L - \alpha_T\right) \omega \cdot \frac{\vec{u} \otimes \vec{u}}{|u|} \tag{1.9}$$

où α_L et α_T sont les dispersivités longitudinales et transversales du milieu.

Comme la perméabilité, la dispersivité n'est pas une propriété locale du milieu et découle d'un changement d'échelle entre celle du pore et celle du VER.

1.1.1.5. Composition minéralogique : nm_i (mol)

La composition minéralogique d'un milieu poreux de volume V_{total} (m^3) est donnée par le nombre de moles nm_i (mol) de chaque minéral qui le compose.

Lorsque ce nombre de moles est rapporté au nombre de moles total, on parle alors de fraction molaire : χMm_i (-) donnée par :

$$\chi Mm_i = \frac{nm_i}{\sum_j nm_j} \tag{1.10}$$

Afin d'assurer la cohérence avec les unités utilisées en chimie des solutions, on peut exprimer la composition minéralogique sous forme de concentration, $\left[cm_i\right]$ (mol.m^{-3}) , en moles de chaque minéral par volume total du milieu [124] :

$$[cm_i] = \frac{nm_i}{V_{total}} \tag{1.11}$$

Elle peut également être donnée en fraction volumique, χVm_i (-) définie comme le ratio entre le volume occupé par le minéral *i* : Vm_i et le volume total de la phase solide :

$$\chi Vm_i = \frac{Vm_i}{\omega \cdot V_{total}} = \frac{V_{mol,i} \cdot nm_i}{\omega \cdot V_{total}} \tag{1.12}$$

Cela peut également être exprimé en fonction du nombre de moles nm_i en faisant intervenir le volume molaire du minéral $V_{mol,i}$ ($m^3.mol^{-1}$).

1.1.1.6. Surface spécifique : A ($m^2.kg^{-1}$)

La surface spécifique d'un milieu poreux se définit comme la surface disponible pour des réactions à l'interface liquide-solide. Elle est traditionnellement ramenée à une masse donnée de solide, exprimée en m^2/g.

Il a été proposé certaines règles, définies par les équations (1.3) à (1.5), pour contraindre les relations porosité-tortuosité. Il semble que de telles règles n'aient pas été explicitées pour des relations porosité-surface spécifique. Nous proposons les règles suivantes :

(i) En absence de milieu poreux, la surface spécifique est nulle :

$$\lim_{\omega \to 1} A = 0 \tag{1.13}$$

Il convient de souligner que cette limite a un caractère extrêmement théorique. En effet, il nous semble difficile de décrire l'évolution complète d'un milieu poreux ; de son colmatage à sa disparition avec une seule loi. Lorsque la porosité tend vers 1, on assiste à une disparition du squelette solide, les grains ne sont plus en contact les uns avec les autres et le milieu va alors se déformer. Le recours à un autre formalisme plus complexe devient alors évident.

(ii) Lorsque le milieu poreux est totalement colmaté, la surface spécifique est nulle :

$$\lim_{\omega \to 0} A = 0 \tag{1.14}$$

Il est possible de mettre en avant la complexité des relations entre porosité et surface spécifique en se penchant sur le cas très simple d'un milieu poreux constitué de sphères de rayon uniforme R. Dans ce cas, la porosité d'un tel milieu dépend entre autre du type d'arrangement (modèle cristallin) entre les sphères et varie entre 0,476 dans le cas d'un système cubique à 0,26 pour un système rhomboédrique. Si l'on considère une masse M de ce milieu poreux, en notant ρ_S la masse volumique du solide, le nombre de sphères n contenues dans cette masse est :

$$n = \frac{M}{\rho_S \frac{4}{3}\pi R^3} \tag{1.15}$$

La surface spécifique de ce milieu poreux est alors, quel que soit le modèle cristallin :

$$A = \frac{n \cdot 4\pi R^2}{M} = \frac{3}{\rho_S R} \tag{1.16}$$

Ainsi, cet exemple très simple montre que la construction de relations entre porosité et surface spécifique ne sera, au mieux, possible que dans des conditions très strictes en fortement dépendantes des conditions expérimentales.

1.1.1.7. Capacité calorifique : Cp ($J.kg^{-1}.K^{-1}$)

La capacité calorifique Cp représente l'apport d'énergie nécessaire à l'élévation de température du milieu. Elle dépend de la composition de celui-ci. C'est une grandeur thermodynamique intensive. Suivant les approches, on parle de capacité calorifique massique, volumique ou molaire.

Le Tableau 1-1 montre que les capacités calorifiques des roches sont souvent très proches, et plutôt faible par rapport à celle de l'eau : 4 184 $J.kg^{-1}.K^{-1}$.

Tableau 1-1 : Capacité calorifique de certaines roches.
(source : Homand [125])

Roche	Capacité calorifique massique $J.kg^{-1}.K^{-1}$
Granite (Limousin)	700
Ardoise (Angers)	740
Argilite (Tournemire)	815
Marne (Alsace)	826
Argilite (Aisne)	845
Calcaire (Euville)	846
NaCl	870

1.1.1.8. Conductivité thermique λ_T ($W.m^{-3}.K^{-1}$)

La conductivité thermique représente la capacité du milieu à conduire la chaleur. Elle dépend de sa composition, mais également de la géométrie porale du milieu. Elle varie selon l'anisotropie du milieu, elle est plus forte selon les plans principaux, plus faible perpendiculairement à ceux-ci. Ce point amène à suggérer l'usage d'un tenseur de conductivité thermique. Le Tableau 1-2 montre la conductivité thermique de certaines roches, et met en évidence l'importance de l'anisotropie en donnant les valeurs selon l'axe principal de la roche et perpendiculairement à celui-ci.

Tableau 1-2 : Conductivité thermique de certaines roches.
(source : Homand [125])

Roche	Conductivité thermique $W.m^{-3}.K^{-1}$
Grès (Vosges)	2,7
Granite (Limousin)	2,8
Ardoise (Angers)	1,2 et 4,5
Argilite (Tournemire)	0,7 et 2
Marne (Alsace)	1,04 et 1,4
Argilite (Aisne)	0,75 et 1,4
Calcaire (Euville)	3,5
NaCl	6

La conductivité thermique telle que nous l'avons présentée n'est pas une propriété locale du milieu et découle d'un changement d'échelle entre celle du pore et celle du VER. Cependant, la plage de variation de la conductivité (de 1 à 6 $W.m^{-1}.K^{-1}$) est relativement limitée, alors que la perméabilité varie sur 7 à 8 ordres de grandeur. De plus, en milieu poreux saturé, la conductivité apparente du milieu est la résultante de celle de la roche et de celle de l'eau. Or la conductivité thermique de l'eau (0,60 $W.m^{-1}.K^{-1}$) est relativement proche de celle des roches. Ainsi, pour la plupart des roches, et quelle que soit la configuration géométrique de la porosité locale, la conductivité thermique apparente se situera dans une plage assez restreinte : entre 0,60 et environs 3 ou 4 $W.m^{-1}.K^{-1}$.

1.1.2. EVOLUTION DU MILIEU

Lorsque des phénomènes chimiques ont lieu au sein d'un milieu poreux, la structure de celui-ci peut être modifiée [3]. Cela peut se produire entre autre lors :

- des réactions de précipitation et de dissolution, qui entrainent la croissance ou la décroissance, l'apparition ou la disparition de phases minérales. Ces phénomènes sont très largement abordés lors de l'étude de la diagénèse [40, 74, 120].
- des phénomènes de colmatage, liés à la croissance de microorganismes où à la mobilisation / redéposition de colloïdes ou de nanoparticules. A l'heure actuelle, ces phénomènes sont relativement peu abordés dans la littérature dédiée à la modélisation numérique et les travaux qui s'y rapportent sont extrêmement récents.

L'évolution de la structure du milieu poreux est habituellement décrite comme la propagation en cascade des modifications d'une propriété sur l'autre. Les phénomènes chimiques (précipitation-dissolution de minéraux, relargage ou capture de particules fines, échange d'ions et gonflement des argiles) entraînent une modification de la composition du milieu, représentée fréquemment par des variations de fraction volumique de chaque minéral. Des lois sont ensuite développées pour lier les fractions volumiques des minéraux à la porosité totale du milieu. Enfin, la porosité totale du milieu permet, sous certaines hypothèses, de déterminer les autres propriétés du milieu : porosité efficace, tortuosité, perméabilité, surface spécifique… La Figure 1-3 permet de visualiser les interactions entre ces grandeurs et de hiérarchiser les différentes boucles de rétroaction.

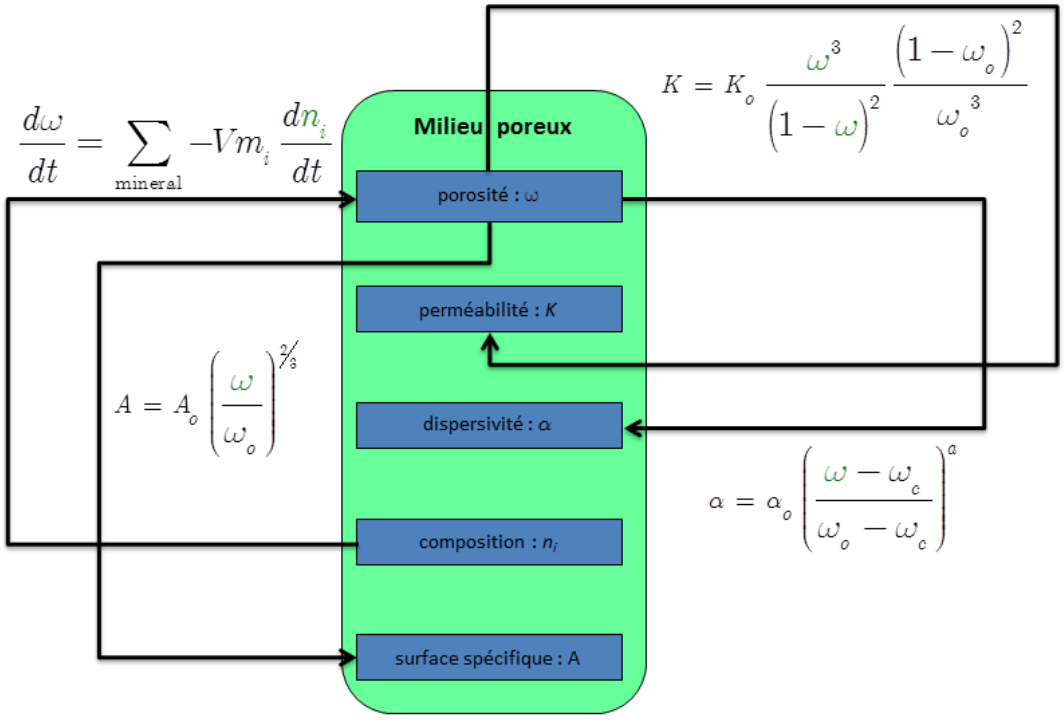


Figure 1-3 : Interactions entre les paramètres décrivant la structure d'un milieu poreux

1.1.2.1. Composition minéralogique

1.1.2.1.1. Porosité

L'évolution des différentes réactions chimiques peut entraîner des modifications de la composition (minéralogique entre autre) du milieu poreux. C'est par le biais de cette modification de composition que les phénomènes chimiques vont induire les changements des différentes propriétés du milieu poreux.

Si on se rapporte à la définition de la porosité, on peut écrire la relation suivante entre porosité et composition minéralogique pour un volume donné de milieu poreux :

$$\omega = 1 - \frac{\sum_i Vm_i}{V_{total}} = 1 - \frac{\sum_i V_{mol,i} \cdot nm_i}{V_{total}} \tag{1.17}$$

Exprimée en terme de concentration, cette relation s'écrit [45] :

$$\omega = 1 - \sum_{i,mineral} V_{mol,i} \cdot [cm_i] \tag{1.18}$$

Historiquement, la variation de la porosité au cours du temps a été écrite [40, 109, 126] en fonction des vitesses de réactions de précipitation-dissolution rm_i :

$$\frac{d\omega}{dt} = -\sum V_{mol,i} \cdot rm_i \tag{1.19}$$

Cependant, cette formulation suppose une formulation cinétique des réactions de précipitation dissolution afin de disposer des vitesses de réaction. Si on choisit une formulation à l'équilibre instantané, on utilisera plutôt la relation suivante, obtenue par dérivation temporelle de l'équation (1.17) :

$$\frac{d\omega}{dt} = -\frac{1}{V_{total}} \sum_i V_{mol,i} \cdot \frac{d\left(nm_i\right)}{dt} \tag{1.20}$$

Si on se rapporte à une formulation en concentration de la composition minéralogique (1.11), on a la formulation suivante :

$$\frac{d\omega}{dt} = -\frac{1}{V_{total}} \sum_i V_{mol,i} \cdot \frac{d\left([cm_i]V_{total}\right)}{dt} \tag{1.21}$$

Si l'on suppose le volume total du milieu inchangé (pas de déformation mécanique), on obtient :

$$\frac{d\omega}{dt} = -\sum_i V_{mol,i} \cdot \frac{d[cm_i]}{dt} \tag{1.22}$$

Il est à souligner que cette formulation, opérationnelle, ne permet pas de faire la distinction entre les différents types de porosité (connectée ou non, voir Bear [118]). De plus, certains travaux montrent que cette approche ne peut être conservée lorsque la porosité devient trop faible. En effet, les réactions de précipitation et dissolution à l'origine des modifications du volume de la phase solide sont, dans la formulation présentée ici, traitées à pression et température constantes, et leurs vitesses sont contrôlées par les paramètres physico-chimiques classiques : température, pression, surface spécifique, concentration des réactifs et des catalyseurs. Cependant, lorsque la porosité diminue, Merino et Dewers [127] soulignent que ce n'est plus la pression qui est constante, mais le volume du solide. Dans ce cas, la formation d'une quantité donnée de solide engendrerait une contrainte mécanique augmentant ainsi la pression et générant des déformations du milieu.

1.1.2.1.2. Capacité calorifique

Pour chaque minéral, la capacité calorifique change selon la température. On trouve [128] la relation empirique suivante donnant, la capacité calorifique sur une large plage de température :

$$Cp_i(T) = a + b \times T + c \times T^2 + \frac{d}{T^2} \tag{1.23}$$

avec *a*, *b*, *c* et *d* des coefficients dépendant du matériau considéré.

Pour un milieu poreux de porosité ω et de composition molaire n_i, la capacité calorifique massique du solide s'exprime :

$$Cp_s = (1-\omega)\sum_i \chi Mm_i \cdot M_i \cdot Cp_i \tag{1.24}$$

où Cp_i est la capacité calorifique massique de chacun des minéraux composant le milieu et M_i leurs masses molaires respectives.

1.1.2.2. La porosité pour déterminer les autres propriétés

La porosité est très souvent utilisée comme paramètre clé pour contrôler l'évolution des autres propriétés du milieu poreux : perméabilité, diffusion effective, surface spécifique… Les rétroactions mises en évidence dans la Figure 1-3 explicitent ce choix.

1.1.2.2.1. Tortuosité

Il existe de nombreuses relations, théoriques ou empiriques, permettant d'obtenir la tortuosité à partir d'une valeur de porosité. Des compilations de ces relations sont données par différents auteurs [121, 122].

Tableau 1-3 : Compilation des relations porosité-tortuosité
(source Boudreau [122] et Shen et Chen [121])
Les chiffres (1) et lettre (B-F) entre parenthèse permettent de repérer les courbes sur la Figure 1-4

Relation théorique	Relation empirique Les valeurs des paramètres sont les meilleures estimations fournies par Boudreau sur un ensemble de données expérimentales
(1) $\tau^2 = \frac{(3-\omega)}{2}$	(A) $\tau^2 = a\omega^{1-m}$ Loi de Archie $a = 1$ et $m = 2{,}14 \pm 0{,}03$
(4) $\tau^2 = 2-\omega$	
(5) $\tau^2 = \omega^{-1/2}$	
(2) $\tau^2 = \omega^{-1/3}$	(B-F) $\tau^2 = \omega + a(1-\omega)$ Equation de Burger-Frieke $a = 3{,}79 \pm 0{,}11$
(3) $\tau^2 = 1 - \frac{1}{2} ln(\omega)$	
(8) $\tau^2 = 1 - ln(\omega)$	
(6) $\tau^2 = \frac{\omega}{1-(1-\omega)^{1/3}}$	(M-W) $\tau^2 = 1 - a \cdot ln(\omega)$ Relation de Weissberg modifiée $a = 2{,}02 \pm 0{,}08$
(7) $\tau^2 = \left(\frac{2-\omega}{\omega}\right)^2$	

Cependant, la grande variabilité de structure des milieux poreux naturels ne permet pas de dégager une relation universelle simple (ou non) comme cela peut se voir sur la Figure 1-4 qui superpose les relations théoriques présentées dans le Tableau 1-3 et une compilation de valeurs expérimentales de porosité et tortuosité.

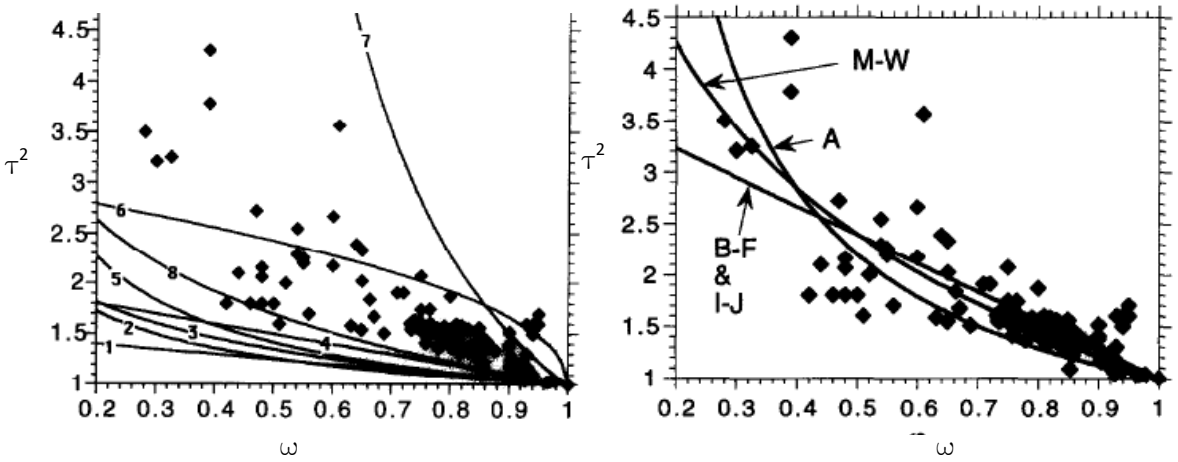


Figure 1-4 : Relation porosité-tortuosité, résultats expérimentaux et courbes théoriques
(source : Boudreau 1996 [122])
Les équations théoriques sont présentées dans le Tableau 1-3

Il semble néanmoins, comme cela se remarque sur la Figure 1-4, que les modèles théoriques ont pour une grande part (modèles 1, 2, 3, 4, 5 et 8) tendance à fournir une tortuosité plus faible que les valeurs expérimentales. Le modèle 6 présente également cette tendance à la sous-estimation pour les porosités faibles. Le modèle 7 propose plutôt une très nette surestimation. Les modèles empiriques, une fois le jeu de paramètres correctement calé, fournissent des valeurs de tortuosités passant par le nuage de point. Cependant, la dispersion des valeurs expérimentales (ou la grande variété des milieux poreux naturels), en permet pas de proposer un modèle unique.

1.1.2.2.2. Perméabilité :

De nombreux modèles existent pour lier porosité et perméabilité. Les premières relations remontent à la formule de Hazen, liant la perméabilité au diamètre moyen des grains. Cependant, cette relation est souvent considérée comme trop simple pour décrire un milieu poreux naturel [129].

Le modèle Kozeny-Carman [118, 130] permet de déterminer la perméabilité sans tenir compte de l'état initial. Cependant, il fait appel à davantage de paramètres décrivant le milieu, incluant la la surface spécifique A, la tortuosité et c_{K-C} le paramètre de forme du milieu :

$$\mathbf{k} = \frac{1}{c_{K-C} \cdot \tau A^2} \frac{\omega^3}{(1-\omega)^2} \tag{1.25}$$

Ainsi, l'utilisation de ce modèle pour décrire l'évolution d'un milieu poreux impose de développer des lois donnant l'évolution de la tortuosité et de la surface spécifique.

Certains modèles de transport réactif utilisent une forme dérivée de la relation de Kozeny-Carman [73, 131] donnant la perméabilité pour une valeur de porosité donnée en fonction de la perméabilité $\mathbf{k_0}$ et de la porosité ω_0 du milieu à l'instant initial :

$$\mathbf{k} = \mathbf{k_0} \frac{\omega^3}{(1-\omega)^2} \frac{(1-\omega_0)^2}{{\omega_0}^3} \tag{1.26}$$

Cette relation est obtenue en divisant la relation (1.25) à une valeur de porosité donnée par la même relation à une valeur de porosité de référence ω_0. L'intérêt de ce modèle est de ne faire appel qu'à la porosité et à la perméabilité initiale, ce qui limite l'usage de multiples paramètres parfois difficiles à obtenir. De plus, ce modèle est adapté aux milieux anisotropes, car il permet de conserver la dimension tensorielle de la perméabilité. Cependant, il faut ici faire l'hypothèse que ni la tortuosité ni la surface spécifique ne changent durant la variation de porosité.

Le modèle de Brinkmann [48, 74] permet de déterminer la perméabilité à partir de la seule porosité du milieu :

$$\mathbf{k} = -\frac{1}{18}\left[\frac{3}{4}(1-\omega)\right]^{2/3}\left(3+\frac{4}{1-\omega}-3\sqrt{\frac{8}{1-\omega}}\right) \tag{1.27}$$

Il est à noter que ce modèle ne permet pas la prise en compte de milieux anisotropes.

Le modèle de Fair-Hatch [71, 74, 118] est un modèle empirique qui fait intervenir une description minéralogique plus précise du milieu poreux en prenant en compte, pour chaque minéral, le ratio entre son volume molaire et sa surface spécifique

$$\mathbf{k} = \frac{\omega^3}{\tau(1-\omega)^2\left(\sum_{i,mineral}\frac{\chi Vm_i}{A_i}\right)} \tag{1.28}$$

Où χVm_i est la fraction volumique et A_i la surface spécifique de chaque minéral.

Des travaux récents, basés sur une modélisation du transport réactif à l'échelle du pore [132] ou sur des réseaux de pores artificiels [133] montrent que l'obtention de relation porosité-perméabilité est un exercice extrêmement délicat et dépendant à la fois du milieu, de l'hydrodynamique et de la plage de variation de porosité envisagé. Une compilation de mesures expérimentales de porosité et perméabilité sur de nombreux milieux poreux montre (voir Figure 1-5) l'amplitude des variations possibles. Zinszner [134] utilise des relations porosité-perméabilité basées sur une expression polynomiale en $log(k)$ en fonction de $log(\omega)$. Selon la Figure 1-5, l'ordre du polynôme varie entre 3 et 7.

$$log(k) = \sum_{i=0}^{n} a_i \cdot \left(log(\omega)\right)^i \tag{1.29}$$

où n est l'ordre du polynôme et a_i sont les coefficients d'ordre i.

Cependant, malgré la difficulté de la tâche, ce point reste prépondérant dans la modélisation de systèmes hydro-géo-chimiques en évolution. Ainsi que le soulignent Saripalli *et al.* [3] : *il est clair que les modèles reliant les réactions aux modifications spatiales et temporelles de la porosité et de la perméabilité sont un composant essentiel lors de la modélisation.*

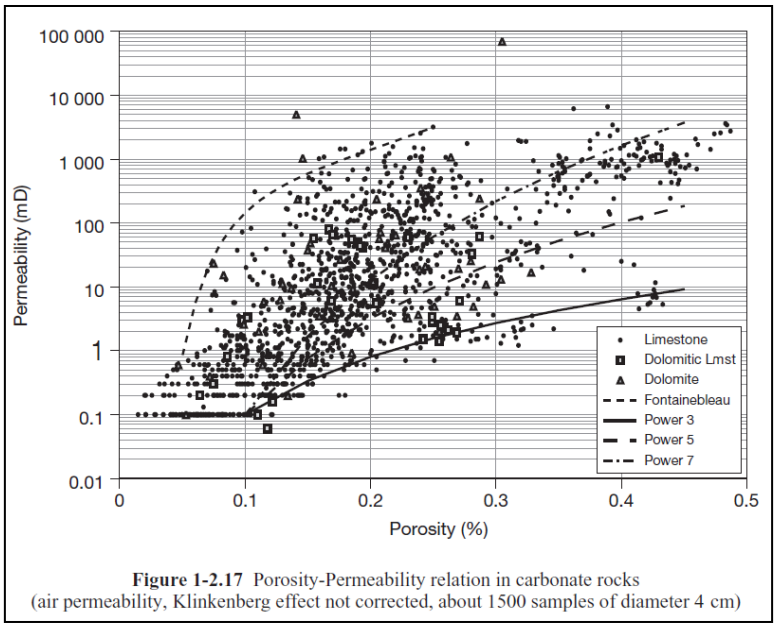


Figure 1-2.17 Porosity-Permeability relation in carbonate rocks
(air permeability, Klinkenberg effect not corrected, about 1500 samples of diameter 4 cm)

Figure 1-5 : Relation porosité-perméabilité dans des roches carbonatées
(source : Zinszner 2007 [134])

1.1.2.2.3. Diffusion moléculaire et dispersivité

La diffusion moléculaire d'un soluté dans un milieu poreux, d_i, n'est pas identique à sa diffusion en solution libre $d_{i,w}$. La diffusion en milieu poreux dépend de la tortuosité du milieu. Avec la définition de la tortuosité donnée par l'équation (1.2), la relation liant ces coefficients est donnée par [121, 122] :

$$d_i = \frac{d_{i,w}}{\tau^2} \tag{1.30}$$

Par le biais d'une relation porosité-tortuosité (voir Tableau 1-3) on peut à partir d'une valeur de porosité obtenir la diffusion moléculaire en milieu poreux.

Certains auteurs [74] introduisent une forme d'anisotropie dans la diffusion moléculaire en fonction de celle du milieu poreux en construisant un tenseur de diffusion $\mathbf{d}_i$:

$$\mathbf{d}_i = \begin{bmatrix} \frac{d_{i,w}}{\tau_x} & & \\ & \frac{d_{i,w}}{\tau_y} & \\ & & \frac{d_{i,w}}{\tau_z} \end{bmatrix} \tag{1.31}$$

Le lien entre la diffusion moléculaire d'un milieu poreux et sa porosité est souvent [71, 131] décrit en utilisant la loi de Archie (voir Tableau 1-3), ce qui donne :

$$d_i = d_{i,w}\omega^{m-1} \tag{1.32}$$

Il est possible d'introduire l'influence de la température dans la détermination de la diffusion moléculaire [135]. La relation proposée combine une loi similaire à celle de Archie et un formalisme proche d'une loi d'Arrhénius pour la prise en compte la température :

$$d_i = d_{i,w} \cdot exp\left[-\frac{E_a}{R}\left(\frac{1}{T}-\frac{1}{T_0}\right)\right]\cdot\omega^m \tag{1.33}$$

où T_0 est la température de référence à laquelle est donnée la diffusion libre (25°C) et E_a l'énergie d'activation du processus de diffusion.

Il a été montré [136] que la dispersivité d'un milieu poreux change relativement peu, même si la porosité évolue et de nombreux auteurs ont négligé les modifications de dispersivité induites par les changements de porosité [131], comme cela a été le cas de la plupart des travaux menés en modélisation des transferts réactifs en milieu poreux. Ceci s'explique, entre autre, par des travaux menés en situation purement diffusive [137].

Il existe néanmoins quelques pistes permettant de construire des relations entre la dispersivité d'un milieu poreux et ses autres propriétés. Bear [118] présente quelques travaux qui tendent à montrer un lien fort entre la structure d'un milieu poreux et sa dispersivité :

Si l'on assimile le milieu poreux à un ensemble de tubes capillaires de rayon R, on peut montrer que l'écoulement laminaire du fluide conduit à une dispersion mécanique des solutés et que cette dispersion mécanique dépend du rayon des capillaires, de la vitesse moyenne de l'écoulement U et de la diffusion moléculaire du soluté :

$$D_m = \frac{R^2 \cdot U^2}{48 d_{i,w}} \tag{1.34}$$

Comme il est également possible de construire des lois donnant la perméabilité d'un réseau de capillaires en fonction de leur rayon, on voit immédiatement qu'un lien fort existe entre ces deux grandeurs.

Mais il n'existe que très peu de travaux étudiant ce lien entre perméabilité et dispersivité. La seule étude trouvée est due à Harleman *et al.* [138]. Ces auteurs étudient expérimentalement le lien entre dispersivité transversale et perméabilité pour différents sables et pour des milieux synthétiques à base de sphères. Ces auteurs montrent que l'on peut déterminer trois paramètres a, b et m tels que :

$$\frac{\rho_f \cdot U \cdot \alpha_L}{\mu_f} = a\left(\frac{\rho_f \cdot U \cdot d_{50}}{\mu_f}\right)^m = b\left(\frac{\rho_f \cdot U \cdot \sqrt{k}}{\mu_f}\right)^m \tag{1.35}$$

Ils montrent que, pour tous les milieux étudiés, le paramètre m est identique, égal à 1,2 (voir Figure 1-6). Pour un milieu composé de grain sphérique, ils donnent a = 0,66 et b = 54 ; et pour des sables, ils donnent a = 0,90 et b = 88. On obtient bien une relation entre dispersivité et perméabilité, mais comme le coefficient m n'est pas égal à 1, cette relation dépend de la vitesse du fluide :

$$\alpha_L = b\left(\frac{\rho_f \cdot U}{\mu_f}\right)^{m-1} k^{m/2} \tag{1.36}$$

D'un point de vue théorique, la relation (1.36) n'est pas entièrement satisfaisante, car perméabilité et dispersivité sont normalement des caractéristiques intrinsèques du milieu poreux. On devrait donc pouvoir s'affranchir de la vitesse dans cette relation.

Une autre approche a été réalisée par Sallé *et al.* [139]. Ces auteurs ont générés différents milieux poreux artificiels (plans, réseaux de tubes, fractals) et ont résolus une équation de Stokes et de convection-diffusion sur ces milieux. L'analyse des moments du panache de concentration permet alors de déterminer la dispersion associée. Ces auteurs montrent que la dispersion suit une loi en puissance selon le nombre de Péclet UL/d :

$$\frac{\alpha_L U}{d_i} = \left(\frac{UL}{d_i}\right)^{\alpha_D} \tag{1.37}$$

où L est une longueur caractéristique de l'écoulement et α_D le coefficient de puissance de la loi. En étudiant des milieux poreux de porosités différentes, Sallé *et al.* montrent que la porosité n'a que peu d'influence sur la dispersivité (voir Figure 1-7).

Le coefficient α_D de la loi de puissance dépend de la porosité du milieu, et vaut entre 2 pour un milieu colmaté, environ 1 pour une porosité proche de 0,8 et entre 1,5 et 1 pour un milieu totalement ouvert (Figure 1-8).

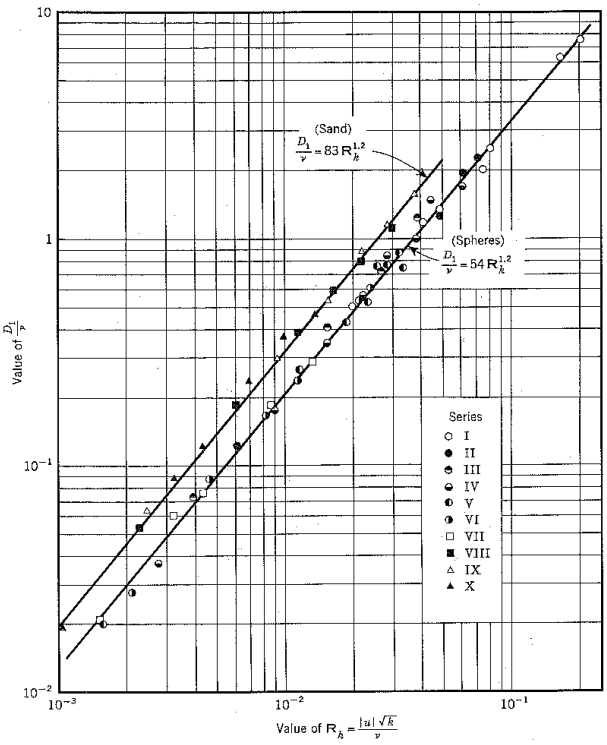


Figure 1-6 : Corrélation expérimentale entre perméabilité et dispersivité
(source : Halreman et al. 1963 [138])

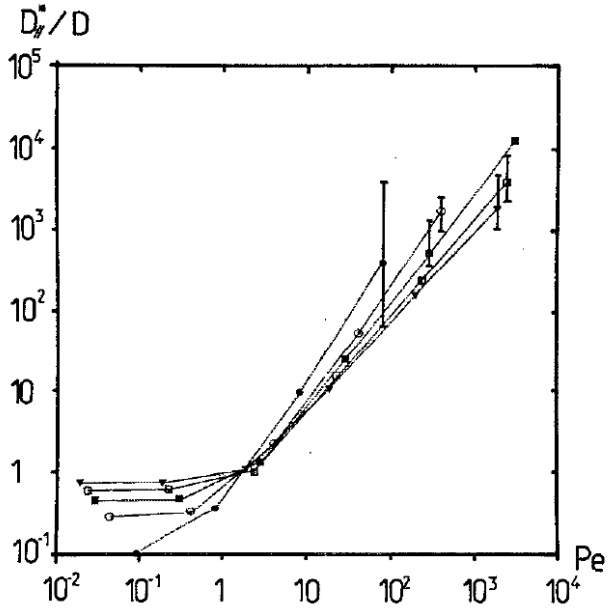


FIG. 16. The dimensionless longitudinal dispersion $D_\parallel^* / D$ in single samples of three-dimensional random media as a function of the Péclet number $\mathrm{Pe} = \xi \bar{v}^* / D$. Data are for: $N_c = 20$; $\epsilon = 0.4$ (●), 0.5 (○), 0.6 (■), 0.7 (□), 0.8 (▼). The intervals of statistical fluctuations for the samples $N_c = 10$ (cf. Table VIII) are indicated by the vertical bars.

Figure 1-7 : Evolution de la dispersivité
en fonction du nombre de Péclet pour différentes porosités.
(Source Sallé et al. 1992 [139])

En général, les travaux de modélisation qui prennent en compte la modification de la structure du milieu poreux négligent les variations de la dispersivité. C'est par le biais de la modification de la vitesse d'écoulement que le tenseur de dispersion évolue avec la structure du milieu. Cette approximation peut en effet se justifier car, comme le montre la Figure 1-7, la sensibilité de la dispersivité au regard de la porosité est beaucoup plus faible que celle par rapport à la vitesse, donc au nombre de Péclet.

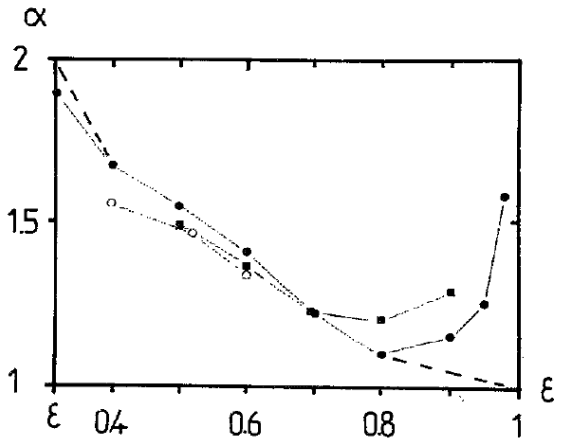


FIG. 17. The exponent α as a function of the porosity ϵ for three-dimensional random media with $N_c=20$ (●) or $N_c=10$ (■). The broken lines are extrapolations toward the asymptotic laws (96) and (98). The open circles (○) correspond to the values of Fontainebleau sandstones (cf. Fig. 20) plotted as a function of $\epsilon+0.3$.

Figure 1-8 : Evolution du coefficient de puissance α_D de la dispersivité en fonction de la porosité pour différents solides.
(Source Sallé et al. 1992 [139])

1.1.2.2.4. Surface spécifique

Les modèles de transport réactif recensés ne font que très peu appel à une quelconque modification de la surface spécifique au cours de l'évolution des systèmes. Les principales approches trouvées sont menées dans le cadre d'un milieu poreux idéal, constitué d'un empilement de sphères [72, 120, 140, 141].

Lichtner [113] ou Emmanuel [72] proposent le modèle suivant, considérant que le milieu poreux est constitué de grains sphériques. :

$$A = A_0 \left(\frac{1-\omega}{1-\omega_0} \right)^{2/3} \tag{1.38}$$

où A_0 est la surface spécifique du milieu poreux à la porosité ω_0. L'analyse de ce modèle montre que, si la condition (1.13) est respectée, ce n'est pas le cas de la condition (1.14) car à porosité nulle, la surface spécifique est ici égale à la surface spécifique initiale A_0. Ce point peut être constaté sur la

Figure 1-9 tirée des travaux de Kieffer *et al.* [141], où l'on remarque nettement que ce modèle est le seul à ne pas tendre vers une surface spécifique nulle à porosité nulle.

En considérant que ce sont les pores du milieu qui sont sphériques, Lichtner [113] ou Merino *et al.* [142] proposent une modification au modèle précédent :

$$A = A_0 \left(\frac{\omega}{\omega_0} \right)^{2/3} \tag{1.39}$$

L'analyse de ce modèle montre un comportement inverse du précédent En effet, la condition (1.14) est ici respectée, alors que ce n'est pas le cas de la condition (1.13). Pour ce modèle, la surface spécifique tend vers $A_0 \cdot \omega_0^{-2/3}$ lorsque la porosité tend vers 1. Cependant, comme nous l'avons souligné précédemment, la description d'un milieu poreux dont la structure disparait imposerait la prise en compte des déformations de celui-ci.

Canals *et al.* [140] proposent un modèle plus complexe. En partant de sphères empilées dans un système cubique faces centrées, ils envisagent la croissance du rayon de ces sphères et l'interpénétration de celles-ci. Ils définissent les relations suivantes donnant le volume de chaque grain V_{grain} ainsi que sa surface de contact avec le fluide A_{grain} :

$$\begin{aligned} V_{grain} &= \frac{4}{3}\pi \left(r_0 + dr \right)^3 - 12\pi \left(\frac{2}{3} dr^3 + r_0 dr^2 \right) \\ A_{grain} &= 4\pi \left(r_0 + dr \right)^2 - 24\pi \left(r_0 + dr \right) dr \end{aligned} \tag{1.40}$$

où r_0 est le rayon initial des sphères et dr l'accroissement de ce rayon dû aux réactions chimiques. Ils définissent aussi une valeur d'accroissement critique correspondant à la fermeture des pores et pour laquelle la surface spécifique est alors nulle.

Le nombre de sphère par unité de volume de milieu poreux N_{grain} est supposé constant et est donné en fonction du volume initial des grains $V_{grain,0}$:

$$N_{grain} = \frac{1 - \omega_0}{V_{grain,0}} \tag{1.41}$$

A chaque instant, la porosité et la surface spécifique sont données par :

$$\begin{aligned} \omega &= 1 - N_{grain} \cdot V_{grain} \\ A &= N_{grain} \cdot A_{grain} \end{aligned} \tag{1.42}$$

Afin d'exploiter ce type de modèle d'une façon similaire à ceux présentés précédemment, il faut déterminer une relation entre porosité et surface spécifique.

Cependant, Kieffer *et al.* [141] montrent en comparant des résultats expérimentaux obtenus sur du sable de Fontainebleau et ces trois modèles, que la correspondance est loin d'être parfaite. Néanmoins, l'évolution générale de la surface spécifique en fonction de la porosité est correctement prévue par les modèles décrits par les équations (1.39) et (1.40) à (1.42) et les valeurs sont correctes à un facteur inférieur à 2. En revanche, le modèle donné par l'équation (1.38) ne correspond pas aux résultats expérimentaux.

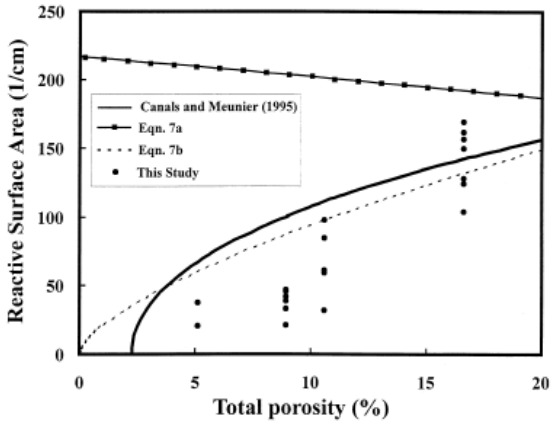


Figure 1-9 : Surface spécifique du sable de Fontainebleau en fonction de la porosité.
Les points correspondent aux valeurs expérimentales ; Eqn 7a correspond à l'équation (1.38) ; Eqn 7b correspond à l'équation (1.39) ; Le modèle de Canals et Meunier est présenté dans les équations (1.40) à (1.42).
(source : Kieffer *et al.* 1999 [141])

1.1.2.2.5. Conductivité thermique $\lambda_{T,eq}$

Il existe plusieurs modèles donnant la conductivité de la matrice poreuse en fonction de sa structure (porosité). De nombreux travaux existent également à l'échelle du pore [143-152].

Pichler *et al.* [148] proposent plusieurs modèles selon la structure du milieu. Pour des pores sphériques, avec une porosité comprise entre 0,2 et 0,8, le modèle MT donne :

$$\lambda_{T,eq} = \lambda_{T,S} \frac{2\omega}{3-\omega} \tag{1.43}$$

avec $\lambda_{T,S}$ la conductivité thermique du solide plein et $\lambda_{T,eq}$ celle du milieu poreux de porosité ω. Le modèle DS suppose une dilution de pores sphériques sans interactions et donne :

$$\lambda_{T,eq} = \lambda_{T,S} \cdot \omega^{3/2} \tag{1.44}$$

Le modèle SCS est adapté aux matériaux polycristallins sans morphologie spécifique des pores. Ce modèle conduit à une conductivité thermique nulle si la porosité est supérieure à 1/3 ; et n'est donc pas adapté aux milieux fortement poreux :

$$\begin{aligned} \lambda_{T,eq} &= \lambda_{T,S}\,\frac{3\omega-1}{2\omega} \quad && \omega < \frac{1}{3} \\ \lambda_{T,eq} &= 0 && \omega > \frac{1}{3} \end{aligned} \tag{1.45}$$

Le modèle FEA (functional equation approach) suppose un milieu dont la porosité est constituée d'une distribution de sphères dont les diamètres sont distribués selon des lois spécifiques. On obtient une loi exponentielle :

$$\lambda_{T,eq} = \lambda_{T,S} \cdot exp\left[\frac{\frac{2}{3}(1-\omega)}{\omega}\right] \tag{1.46}$$

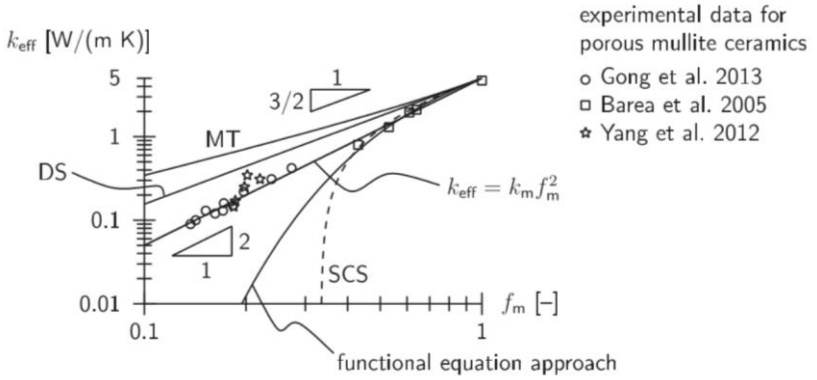


Figure 1-10 : Comparaison des modèles de conductivité thermique.
La porosité ω est notée f_m, la conductivité thermique $\lambda_{T,eq}$ est notée k_{eff}, la conductivité thermique du solide est fixée à $\lambda_{T,S} = k_m = 5$ W/(mK). (Source Pichler *et al.* [148]).

Cependant, l'enjeu de ce travail est nettement moins important que celui relatif à l'évolution de la perméabilité, car comme nous l'avons souligné (§ 1.1.1.8) la conductivité thermique apparente a une plage de variation relativement restreinte (0,6 à 3 ou 4 $W.m^{-1}.K^{-1}$) alors que la perméabilité varie sur 6 à 8 ordres de grandeur.

1.1.3. CONCLUSION MILIEU POREUX

Nous avons montré ici comment, à partir de l'évolution de la composition minéralogie du milieu poreux (représentée dans ce travail par les concentrations des différents minéraux $[cm_i]$), il était possible de décrire les modifications induites sur les autres grandeurs caractéristiques de ce milieu.

La séquence algorithmique serait la suivante : à partir des variations de la composition minéralogiques, on recalcule la porosité du milieu. Cette nouvelle porosité permet d'obtenir de nouvelles valeurs de tortuosité et de surface spécifique. Selon le modèle choisi, on détermine de nouveaux champs de perméabilité et de diffusion moléculaire à l'aide des nouvelles valeurs de porosité, tortuosité et surface spécifique. Enfin, cette nouvelle perméabilité est utilisée pour déterminer une nouvelle dispersivité.

A chaque étape, les verrous sont présents :

Relation composition minéralogique – porosité :

Certes les volumes molaires des différents minéraux purs sont bien connus, mais le cas des solutions solides, assez fréquent en milieu naturel, pose plus de problèmes. De plus, cette relation composition-porosité ne permet pas de déterminer le type de porosité : connectée ou non.

Relations porosité – tortuosité ; surface spécifique ; perméabilité ; diffusion moléculaire

Ces relations sont toutes plus ou moins empiriques, et dépendent fortement du milieu poreux étudié. A l'heure actuelle, les études réalisées portent sur différents types de milieux poreux, mais pas sur un milieu poreux au cours de son évolution.

Relation perméabilité – dispersivité

Nous avons une relation applicable en termes de modélisation, mais qui n'est pas totalement satisfaisante. La relation (1.36) est tout à fait apte à être insérée dans un code de calcul, car la vitesse de l'écoulement par maille ainsi que la perméabilité sont connues à chaque instant, ces paramètres étant utilisés ou fournis par le calcul de l'écoulement de la phase liquide.

On constate que les difficultés sont associées aux grandeurs issues d'un processus de changement d'échelle. Des phénomènes physico-chimiques à l'échelle du pore ne sont pas explicitement décrits à l'échelle du VER et sont approchés par une relation macroscopique. Il est alors extrêmement difficile de fournir des relations macroscopiques valables dans des conditions très diverses, qu'il s'agisse d'utiliser la même relation pour des roches différentes ou pour des conditions d'écoulement différentes, ou dans le cas qui nous intéresse particulièrement, pour décrire une évolution de la nature et de la structure d'un milieu poreux.

1.2. LA PHASE FLUIDE

La phase fluide d'un milieu poreux est le siège d'interactions complexes entres les différentes grandeurs qui la caractérise. Sous l'influence de contraintes extérieures, différents gradients peuvent apparaitre (de charge, de température, de concentration…) qui induiront des déplacements, soit de la phase fluide dans son ensemble (gradient de charge) soit des éléments dissous (gradient de concentration).

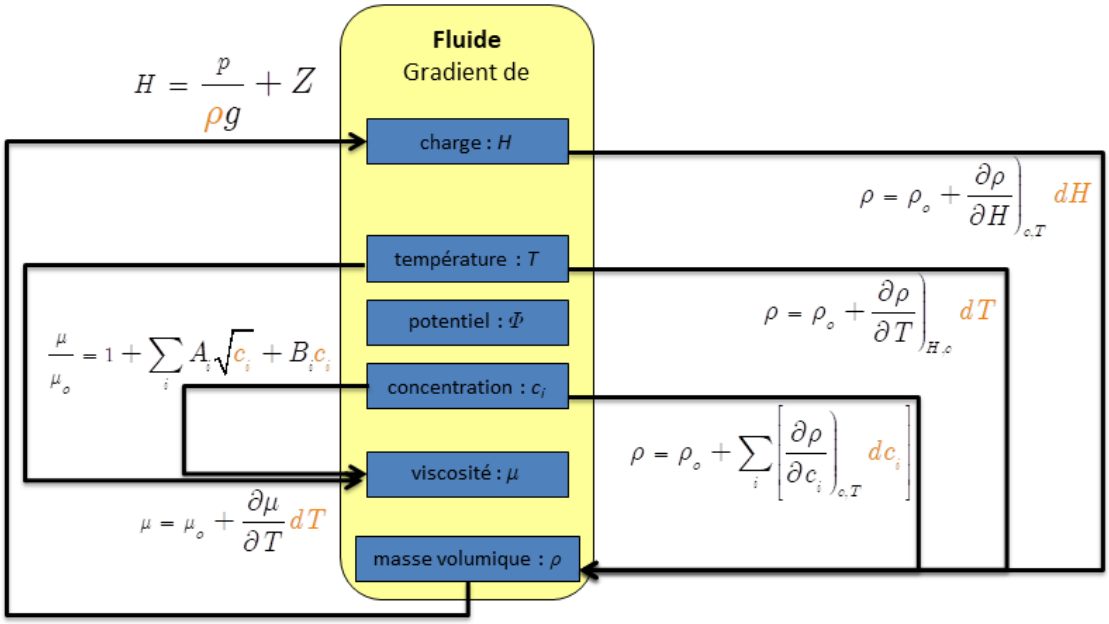


Figure 1-11 : Schéma des interactions au sein de la phase fluide.

1.2.1. DEFINITION DES GRANDEURS

1.2.1.1. Masse volumique : ρ_f ($kg.m^{-3}$)

La masse volumique de la phase fluide est définie comme le rapport de la masse d'une quantité de fluide par son volume.

1.2.1.2. Viscosité μ_f ($kg.m^{-1}.s^{-1}$)

La viscosité dynamique du fluide est définie comme le rapport entre la contrainte tangentielle et la surface de contact.

1.2.1.3. Concentration $[c_i]$ (mol.m^{-3})

Pour chaque élément dissous dans la phase fluide, on définit sa concentration (concentration molaire) comme le rapport entre le nombre de mole du composé dissous et le volume de fluide. C'est la grandeur qui est usuellement utilisée pour décrire les phénomènes chimiques au sein de la phase fluide.

On exprime parfois la quantité de matière en concentration massique $\left[c_{M,i}\right]$ (kg.m^{-3}) :

$$\left[c_{M,i}\right] = M_{mol,i} \cdot \left[c_i\right] \tag{1.47}$$

où $M_{mol,i}$ (kg.mol^{-1}) est la masse molaire du composé dissous.

Cependant, exprimer la réactivité d'un composé en fonction de sa concentration suppose l'hypothèse d'une solution diluée. En introduisant des coefficients de correction d'activité, il est possible de travailler dans ce formalisme jusqu'à des concentrations proche d'une mole par litre. Ceci est donc largement suffisant pour de nombreuses études environnementales. Néanmoins, il est possible d'avoir besoin d'un formalisme plus général, et on utilise alors la fraction molaire χ_i (-) d'un composé :

$$\chi_i = \frac{n_i}{n_w + \sum_k n_k} \tag{1.48}$$

où n_i est le nombre de moles du composé i et n_w est le nombre de moles d'eau. En faisant l'hypothèse d'une solution diluée, c'est-à-dire que la quantité d'eau est très grande devant la quantité des autres composés, on obtient la relation suivante :

$$\chi_i = \frac{n_i}{n_w} \tag{1.49}$$

Ce qui donne, pour les solutions diluées, la relation entre concentration et fraction molaire :

$$\chi_i = V_{mol,w} \cdot \left[c_i\right] \tag{1.50}$$

où $V_{mol,w}$ (m^3.mol^{-1}) est le volume molaire de l'eau.

1.2.1.4. *Capacité calorifique* Cp_w (J.kg^{-1}.K^{-1})

La capacité calorifique de l'eau est très grande, 4 189,3 J/kg/K à 12°C, par rapport à celle des roches (voir Tableau 1-1). Cette valeur change [128] avec la température passant de 4 217,7 J/kg/K à 0°C à une valeur minimale de 4 178,2 J/kg/K à 35°C pour remonter à 4 216,0 J/kg/K à 100°C.

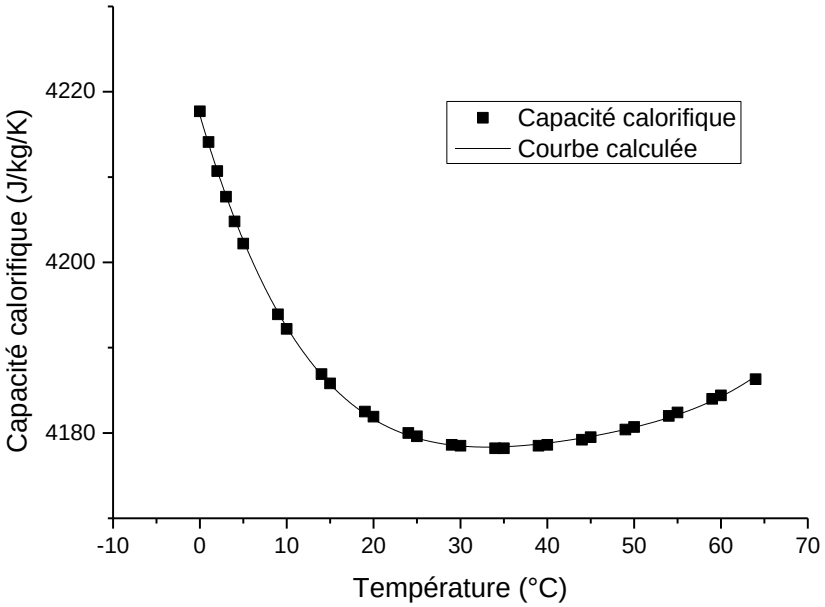


Figure 1-12 : Capacité calorifique de l'eau en fonction de la température.

1.2.2. INTERDEPENDANCES AU SEIN DE LA PHASE FLUIDE

Les travaux dédiés à l'étude des écoulements en milieu poreux sous l'influence de forts contrastes de densité sont nombreux [153-155], souvent liés aux problématique d'intrusion d'eau salée dans les aquifères côtiers [156]. On peut citer deux articles de synthèses sur cette thématique, l'un centré sur les intrusions d'eau salée [156] ; l'autre sur la problématique générale des phénomènes d'écoulement et transport avec contraste de densité en milieu poreux [157].

1.2.2.1. Modification de la masse volumique

La masse volumique du fluide est une fonction de la température, de la pression et de la composition du fluide. Les premiers travaux sur les transferts d'eau salée en milieu poreux [153] ont exprimé les modifications de masse volumique en terme de mélange entre l'eau douce et l'eau salée. Cependant, l'ensemble des auteurs se réfèrent à un formalisme thermodynamique, tel qu'il est présenté par Diersch et Kolditz [157] par exemple.

$$\rho\left(T,p,[c_i]_{i=1,Nc}\right)=\rho_w\left(T°,p°\right)+\left(\frac{\partial\rho}{\partial T}\right)_{p,[c_i]}\cdot dT+\left(\frac{\partial\rho}{\partial p}\right)_{T,[c_i]}\cdot dp+\sum_{i=1}^{Nc}\left(\left(\frac{\partial\rho}{\partial[c_i]}\right)_{p,T,[c_k]_{k\neq i}}\cdot d[c_i]\right) \quad (1.51)$$

où $T°$ et $p°$ sont les température et pression de références pour lesquelles la masse volumique est connue. La relation (1.51) fait appel à différents coefficients :

Le coefficient de compressibilité du fluide :

$$\gamma_w = \frac{\partial \rho}{\partial p}\Big)_{T,[c_i]} \tag{1.52}$$

Le coefficient d'expansion thermique :

$$\beta_w = -\frac{\partial \rho}{\partial T}\Big)_{p,[c_i]} \tag{1.53}$$

Les coefficients d'expansion de solubilisation de chaque espèce dissoute :

$$\alpha_i = \frac{\partial \rho}{\partial [c_i]}\Big)_{p,T,[c_k]_{k\neq i}} \tag{1.54}$$

Sur des plages de variation du paramètre impliqué (pression (1.52), température (1.53), concentration (1.54)) assez faible, on peut considérer que les coefficients de compressibilité ou d'expansion sont constants. Cependant, sur des variations plus importantes, ce n'est plus le cas. Ainsi, Diersch et Kolditz [157] soulignent que le coefficient d'expansion thermique de l'eau varie de $-0{,}68\cdot10^{-4}$ à $7{,}50\cdot10^{-4}\ K^{-1}$ entre 0°C et 100°C. Afin de décrire de larges plages de variation, ces auteurs reportent l'usage de relations polynomiales d'ordre plus ou moins élevées pour exprimer ces coefficients.

Bien qu'à notre connaissance ces travaux n'aient jamais été appliqués au domaine hydro-géo-chimique, de récents développements en thermodynamique appliqués spécifiquement à l'eau [158] permettent, à partir de relations fondamentales, de déterminer les différentes propriétés thermodynamiques de l'eau (capacité calorifique, masse volumique ou vitesse du son) sur des plages de température et de pression très vastes (de -38°C à 1000°C et de 0,1 Pa à 10^6 Pa). Les relations utilisées sont assez complexes et longues et ne seront pas présentées ici. Cependant, il nous semble que, si la pression et la température varient sur de grandes plages, et que l'on souhaite obtenir une modélisation précise du comportement de l'eau, cette approche soit la seule valide. En effet, on peut constater [128] que les coefficients de compressibilité ou d'expansion thermique de l'eau changent avec la pression et la température.

Pour des formulations plus simples, on peut utiliser les éléments suivants : Le Comité International des Poids et Mesures recommande l'usage de la relation de Tanaka *et al.* [159] pour des plages de température allant de 0°C à 40°C.

$$\rho_w(T) = A_{w,5}\left[1 - \frac{\left(t + A_{w,1}\right)^2\left(t + A_{w,2}\right)}{A_{w,3}\left(t + A_{w,4}\right)}\right] \tag{1.55}$$

Diersch [160] reporte l'utilisation d'une interpolation polynômiale d'ordre 6, donnant un résultat parfait. Cependant, comme nous le montrons sur la Figure 1-13, une interpolation à l'ordre 3 donne déjà de très bons résultats. Les valeurs des coefficients de ces trois relations sont données dans le

Tableau 1-4. Il convient de remarquer, comme cela est montré sur la Figure 1-13, qu'un polynôme d'ordre 3 n'est pas adapté pour couvrir la totalité de la plage de température à cause de l'anomalie à 4°C.

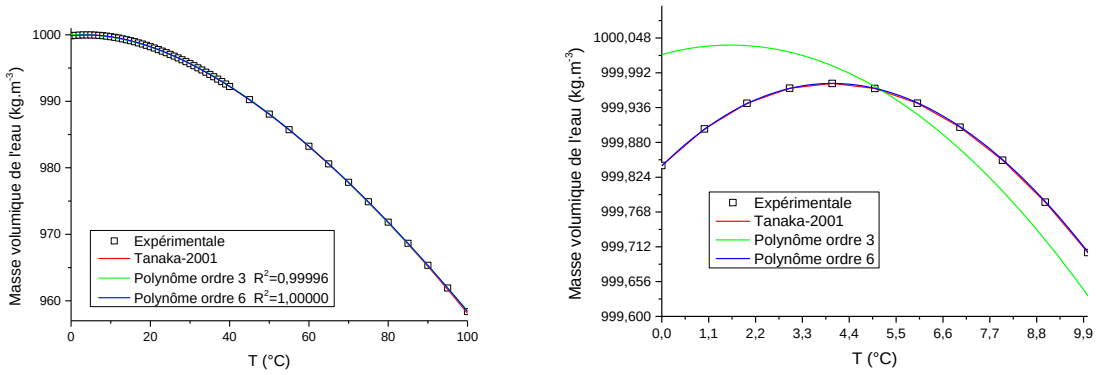


Figure 1-13 : Masse volumique de l'eau en fonction de la température.
Comparaison des données expérimentales à trois modèles Tanaka, polynôme d'ordre 3 et polynôme d'ordre 6
(Source : Valeurs expérimentales d'après [128])

Tableau 1-4 : Valeurs des paramètres des relations température - masse volumique de l'eau.

	Tanaka *et al.* 2001	Polynôme ordre 3	Polynôme ordre 6
$A_{w,0}$		1 000,02132	999,84207
$A_{w,1}$	-3,983035	0,01901	0,0677
$A_{w,2}$	301,797	-0,00594	-0,00897
$A_{w,3}$	522 528,9	1,59988 E-5	8,91831 E-5
$A_{w,4}$	69,34881		-8,20763 E-7
$A_{w,5}$	999,974950		4,44714 E-9
$A_{w,6}$			-1,01513 E-11

Pour inclure l'influence des espèces dissoutes sur la masse volumique de l'eau, Cheng et Yeh [37] utilisent une approche simplifiée, qui néglige les phénomènes de changement de volume liés à la dissolution d'une espèce[1] :

$$\rho_f = \rho_W(T) + \sum_{i,dissou} [c_i] M_i \left(1 - \frac{\rho_W(T)}{\rho_i}\right) \quad (1.56)$$

[1] Lorsque l'on ajoute (par exemple) un acide fort dans de l'eau, on augmente localement les interactions électrostatiques entre les molécules d'eau, ce qui conduit à une diminution de volume : le volume final de l'eau acidifiée est plus faible que la somme du volume d'eau et du volume d'acide.

avec ρ_i la densité intrinsèque de l'espèce i dissoute.

Dans le cadre des phénomènes de transport en milieu poreux, la relation la plus complète est donnée par Diersch [160], relation valable cependant pour de petites plages de variations :

$$\rho_w = \rho_w^0 \left[1 + \gamma_w \left(p - p^0 \right) + \frac{\alpha}{c_s - c^0} \left(c - c^0 \right) - \beta_w \left(T \right) \left(T - T^0 \right) \right] \tag{1.57}$$

où le coefficient d'expansion thermique est calculé en fonction de la température à partir d'un polynôme de degré 6 donné dans le Tableau 1-4. Les limitations de cette relation sont les suivantes :

- Le coefficient de compressibilité isotherme γ_w est constant alors qu'il varie entre 50,88 bar^{-1} à 0°C et 44,15 bar^{-1} à 46°C selon [128]
- Tous les éléments dissouts sont supposés avoir la même influence sur la masse volumique de l'eau, alors que des différences telles que la masse molaire de la molécule ou son ionisation vont jouer un rôle important dans la masse volumique du mélange final.

1.2.2.2. Modification de la viscosité

L'influence des différents paramètres physiques d'un fluide sur sa viscosité sont décrits dans les travaux spécifiques de mécanique des fluides [161]. Il est souvent reporté que la température ou la composition d'un fluide jouent un rôle important, mais que la pression a une influence assez faible.

Température

En mécanique des fluides, l'influence de la température est souvent décrite par la loi empirique d'Andrade [162] :

$$\mu_w \left(T \right) = A_{w,0} \cdot exp \left(\frac{A_{w,1}}{273{,}15 + T} \right) \tag{1.58}$$

Nous obtenons une bien meilleure estimation (voir Figure 1-14) avec la loi d'Andrade modifiée :

$$\mu_w \left(T \right) = A_{w,0} \cdot exp \left(\frac{A_{w,1}}{A_{w,2} + T} \right) \tag{1.59}$$

Diersch utilise la relation suivante [160] :

$$\mu_w \left(T \right) = \frac{\mu_{w,0}}{1 + 0{,}7063 \frac{T - 150}{100} - 0{,}04832 \left(\frac{T - 150}{100} \right)^3} \tag{1.60}$$

On peut noter que cette relation est équivalente à la relation suivante que nous proposons ici, et qui fournit une très bonne estimation :

$$\mu_w(T) = \frac{A_{w,0}}{1 + A_{w,1} \cdot T - A_{w,2} \cdot T^3} \tag{1.61}$$

Il est également possible d'utiliser un modèle polynômial pour lier température et viscosité. On obtient alors une description parfaite pour un polynôme d'ordre 6. Il faut un polynôme d'ordre 5 pour obtenir le même niveau de description que celui fourni par la loi d'Andrade modifiée (R^2 = 0,99998).

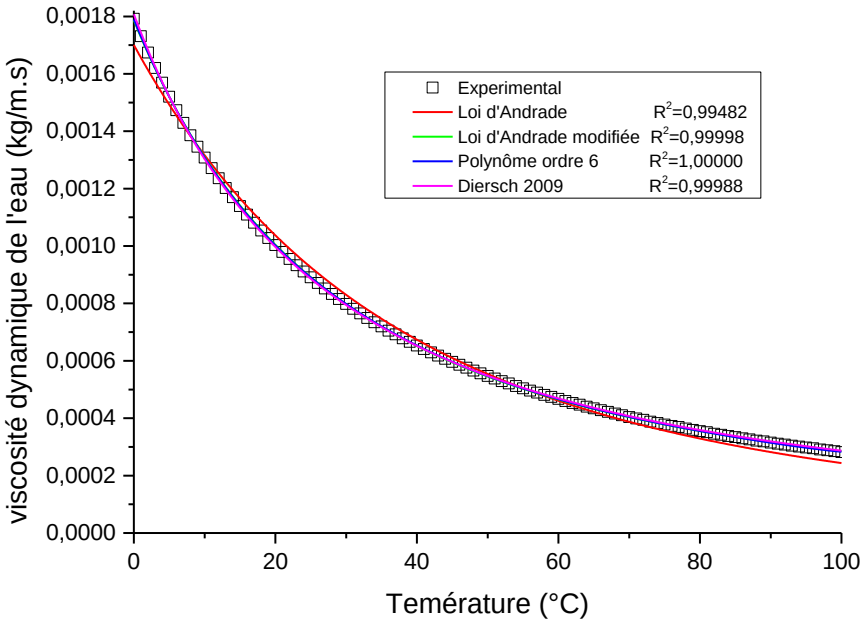


Figure 1-14 : Comparaison des lois empiriques donnant la viscosité dynamique de l'eau en fonction de la température.

Tableau 1-5 : Coefficients des lois de viscosité dynamique de l'eau ($kg.m^{-1}.s$) en fonction de la température (°C)

	Loi Andrade	Andrade modifiée	Diersch modifiée	Polynôme ordre 6
$A_{w,0}$	1,20764E-6	3,1189E-5	0,00183	0,00179
$A_{w,1}$	1980,35821	488,02468	0,04115	-6,05927E-5
$A_{w,2}$		120,56649	1,72431E-6	1,466E-6
$A_{w,3}$				-2,5213E-8
$A_{w,4}$				2,78904E-10
$A_{w,5}$				-1,72863E-12
$A_{w,6}$				4,50121E-15

Composition

Pour prendre en compte l'influence de la compostion chimique, Cheng [37] utilise la relation suivante :

$$\mu_f = \mu_w(T) + \sum_{i,dissout} [c_i]\eta_i \tag{1.62}$$

où η_i est le coefficient d'influence de l'espèce dissoute i sur la viscosité totale. Cette relation à l'avantage d'être assez simple à mettre en œuvre, mais a une base empirique.

La relation de Falkenhagen [161] propose une approche théorique pour des solutions di-ioniques où la valence et la mobilité des 2 ions sont identiques (1.63) :

$$\frac{\mu_f}{\mu_w} = 1 + A\sqrt{[c]} + B_i[c] \tag{1.63}$$

où A est déterminé par (1.64) :

$$A = \frac{0{,}517 \cdot z_i^2}{L_i^\infty \cdot \mu_w \cdot \sqrt{\varepsilon_0} \cdot T} \tag{1.64}$$

avec z_i la charge de la paire d'ions, L_i^∞ sa mobilité à dilution infinie et ε_0 la constante diélectrique de l'eau. Le terme B_i est un terme correcteur qui a été ajouté ultérieurement pour prendre en compte les ions de charges et mobilités différentes.

Tableau 1-6 : Valeurs expérimentales du coefficient B_i dans l'équation (1.63)
(Source Viswanath *et al.* [161])

Cation	*Contribution*	*Cation*	*Contribution*	*Anion*	*Contribution*
Li^+	0.1495	Be^{2+}	0.3923	Cl^-	- 0.0070
Na^+	0.0863	Mg^{2+}	0.3852	Br^-	- 0.0420
K^+	- 0.0070	Ca^{2+}	0.2850	I^-	- 0.0685
Rb^+	- 0.0300	Sr^{2+}	0.2650	OH^-	0.1200
Cs^+	- 0.0450	Ba^{2+}	0.2200	ClO_3^-	0.0240
NH_4^+	- 0.0074	Fe^{2+}	0.4160	NO_3^-	- 0.0460
Ag^+	0.0910	La^{3+}	0.5880	MnO_4^-	- 0.0590
H^+	0.0690	Ce^{3+}	0.5765	SO_4^{2-}	0.2085

Cependant, cette relation demeure restreinte à des solutions di-ioniques, et n'est donc pas simplement applicable à la problématique de nos modèles de transport réactif. Son usage peut s'envisager dans le cas où la phase aqueuse est dominée par un couple d'ions, comme c'est le cas dans des expériences de laboratoire où la force ionique est imposée.

Lencka *et al.* [163] proposent un modèle qui conviendrait mieux à nos besoins. Ces auteurs décrivent la viscosité relative μ_f/μ_w d'une solution aqueuse multi-espèces comme la somme des termes

d'interactions électrostatiques longue distance μ_{LR}, les termes spécifiques à chaque espèce μ_S et des termes d'interactions entre les espèces μ_{S-S}.

$$\frac{\mu_f}{\mu_w} = 1 + \mu_{LR} + \mu_S + \mu_{S-S} \tag{1.65}$$

Le terme d'interactions longue distance est calculé à partir de la théorie développée par Onsager et Fuoss [164] à partir de la force ionique de la solution, de la température et de sa composition chimique :

$$\mu_{LR} = a\frac{1}{\mu_w}\left(\frac{2I}{\varepsilon_w T}\right)^{1/2}\left[\left(\sum_{i=1}^{Nc}\frac{\mu_i z_i}{\lambda_i}\right) - 4\mathbf{r}\sum_{n=0}^{\infty}c_s\mathbf{s}_n\right] \tag{1.66}$$

où $a = 0{,}36454$, ε_w est la constante diélectrique de l'eau, λ_i est la conductivité ionique que l'espèce et les termes μ_i, $\mathbf{r}$ et $\mathbf{s}_n$ se calculent à partir des concentration et des charges des espèces.

Le terme spécifique à chaque espèce exploite l'additivité des coefficients B_i de l'équation (1.63) :

$$\mu_S = \sum_i^{Nc} B_i\left[c_i\right] \tag{1.67}$$

Le terme d'interaction entre les espèces fait intervenir des termes, f_i, dérivés des fractions molaires, en ajustant celles-ci en fonction de la charge des espèces :

$$\mu_{S-S} = \sum_i\sum_j f_i f_j \cdot D_{ij} \cdot I^2 \quad \text{avec} \quad f_i = \frac{\left[c_i\right]/max\left(1,|z_i|\right)}{\sum_k\left[c_k\right]/max\left(1,|z_k|\right)} \tag{1.68}$$

Le détail des relations, parfois lourdes, permettant de calculer les différents paramètres est donné par Lencka *et al.* [163]. Seul le terme D_{ij} contient des éléments empiriques nécessitant un calage expérimental. Le modèle présenté permet également d'inclure l'influence de la température.

On peut remarquer sur la Figure 1-15 que la gamme de concentration couverte par ces relations est extrêmement vaste, beaucoup plus vaste que nécessaire pour un modèle de transport réactif développé sur une chimie exprimée en concentration.

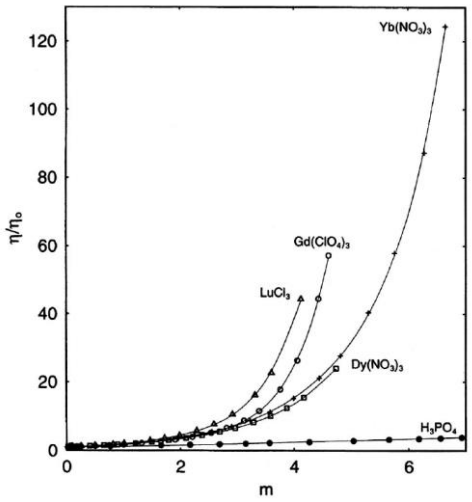


Figure 1-15 : Evolution de la viscosité relative en fonction de la concentration pour différents systèmes chimiques
(source Lencka et al. [163])

Il semble que la relation utilisée par Cheng [37], l'équation (1.62), puisse être assimilée à une simplification de la relation (1.65), dans laquelle seuls les termes spécifiques à chaque espèce μ_S seraient pris en compte.

Ce point peut sembler de faible importance si l'on part du principe que les solutions impliquées dans les codes de transport réactif sont diluées, car les variations de viscosité seraient alors plutôt faibles. Mais si l'on compare les variations de viscosité pour deux solution di-ioniques calculées à l'aide de la relation (1.63), on constate (Figure 1-16) que la présence des solutés peut induire des variations de viscosité relative d'environ 10 %.

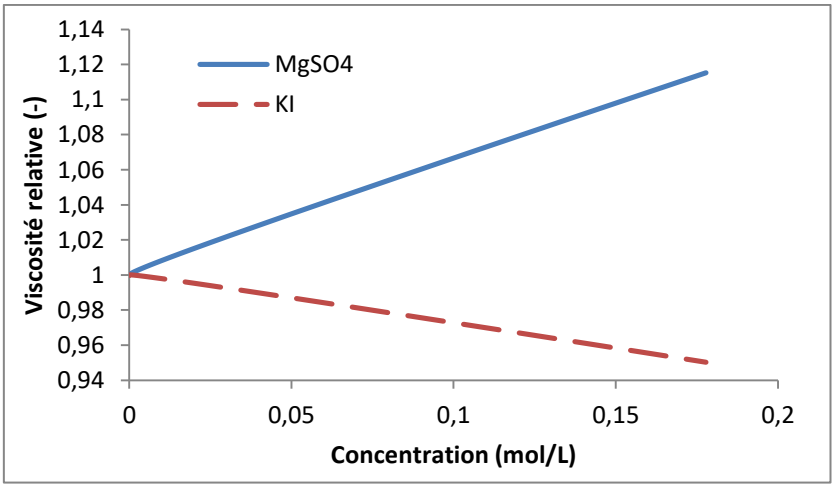


Figure 1-16 : Variations de viscosité relative pour une solution de $MgSO_4$ ou de KI.
(Paramètres A et B selon Viswanath *et al.* [161])

1.2.3. CONCLUSION PHASE FLUIDE

Les relations décrivant l'évolution de la phase fluide ainsi que les différents paramètres impliqués sont plutôt bien connus. Il semble que seule l'influence des différentes concentrations d'un ensemble d'ions sur la viscosité de l'eau soit encore délicate à prendre en compte. En effet, la relation (1.65) proposée par Lencka *et al.* [163] n'est actuellement pas utilisée par les codes de transport réactif, et bien que cette relation soit construite sur une base théorique stable, certains paramètres empiriques demeurent comme les termes B_i de l'équation (1.67) et les paramètres D_{ij} de la relation (1.68). Comme cela peut se remarquer en observant le Tableau 1-6, les valeurs expérimentales de ces coefficients ne sont pas facilement accessibles pour l'ensemble des espèces chimiques : on constate que sont données les valeurs pour les ions *simples* et courants, mais pas pour ceux plus rares ou complexes tels que S^{2-}, CrO_4^{2-} ou $Fe(OH)_2^+$.

1.3. EQUATIONS DE CONSERVATION

Les phénomènes de transport réactif en milieu poreux sont décrits par des équations de conservation.

La conservation de la masse de solide nous donne la relation entre la composition minéralogique et la porosité du milieu, décrite précédemment(1.19).

La conservation de la masse de fluide, associée à la conservation des moments, sont utilisées pour déterminer le champ de vitesse de la phase fluide.

La conservation de la masse de soluté permet de déterminer l'évolution du champ de concentrations au cours du temps.

La conservation de l'énergie décrit le transport de la chaleur au sein du milieu.

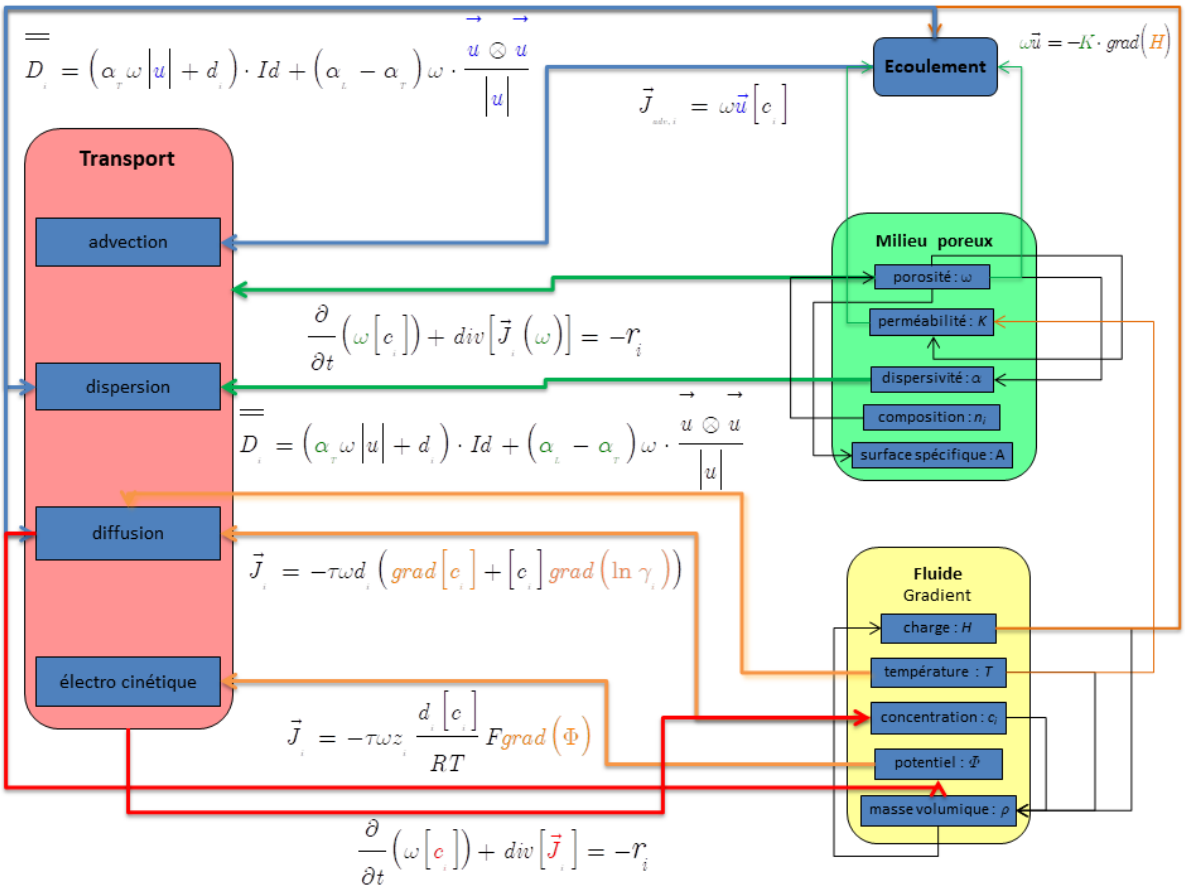


Figure 1-17 : Schéma des interactions liées aux équations de conservation

1.3.1. ECOULEMENT – GRADIENT DE CHARGE

Les relations décrivant l'écoulement du fluide au sein du milieu poreux expriment la conservation des moments de la phase fluide [77]. Suivant le degré de description du milieu poreux, cette relation peut prendre différentes formes.

Avec une description à l'échelle du pore, beaucoup d'auteurs utilisent les équations de Navier-Stockes. Selon la présentation de Kang et Lichtner [16, 165], on a l'équation de continuité :

$$\frac{\partial \rho_f}{\partial t} + \nabla \cdot \left(\rho_f \vec{u} \right) = 0 \tag{1.69}$$

ainsi que l'équation de conservation du moment :

$$\frac{\partial}{\partial t}\left(\rho_f \vec{u} \right) + \nabla \cdot \left(\rho_f \vec{u} \otimes \vec{u} \right) = -\nabla p + \nabla \cdot \sigma + \rho_f g \tag{1.70}$$

où σ est le tenseur des contraintes dues à la viscosité, exprimé de la façon suivante :

$$\sigma = \rho_f \nu \left[\nabla \vec{u} + (\nabla \vec{u})^T - \frac{2}{3}(\nabla \cdot \vec{u})\mathbf{Id} \right] \tag{1.71}$$

A une échelle plus grande, laboratoire ou terrain, cette formulation n'est plus applicable. L'équation de continuité doit faire intervenir la porosité du milieu [118] :

$$\frac{\partial}{\partial t}\left(\omega \rho_f \right) + \nabla \cdot \left(\rho_f \omega \vec{u} \right) = 0 \tag{1.72}$$

et l'écoulement est classiquement décrit par la loi de Darcy [123] :

$$\omega \vec{u} = -K \cdot grad\left(H \right) \tag{1.73}$$

où ω est la porosité, $\vec{u}$ le vecteur vitesse (m.s^{-1}), K est la perméabilité du milieu (m.s^{-1}) et H la charge hydraulique (m). Si l'on explicite la charge hydraulique et la conductivité hydraulique du milieu, on obtient alors :

$$\omega \vec{u} = -\frac{k}{\mu_f} \cdot \left(\nabla p - \rho_f g \right) \tag{1.74}$$

1.3.2. TRANSPORT – GRADIENT DE CONCENTRATION

Pour un soluté réactif c_i, l'équation de conservation nous dit que la quantité de matière consommée par les réactions chimique, $-\sum_{réactions,r} b_{i,r} \cdot r_r$, est égale à la variation de concentration, $\frac{\partial}{\partial t}(\omega[c_i])$, à laquelle on ajoute la matière apportée par le flux physique, $div(\vec{J}_i)$.

$$\frac{\partial}{\partial t}(\omega[c_i]) + div(\vec{J}_i) = -\sum_{réactions,q} b_{i,q} \cdot r_q \tag{1.75}$$

avec $[c_i]$ la concentration molaire du soluté (mol.m^{-3}), $\vec{J}_i$ (mol.m^{-2}.s^{-1}) le flux de matière, $b_{i,q}$ le coefficient stœchiométrique du soluté c_i dans la réaction q, et r_q la vitesse de la réaction q.

Suivant la complexité des phénomènes à prendre en compte, le flux de matière peut s'exprimer de différentes façons. Il est assez courant de prendre la formulation proposée par Lichtner en 1985 [109] :

$$\vec{J}_i = \omega\vec{u}[c_i] - \overline{\overline{D_i}} \cdot grad[c_i] \tag{1.76}$$

avec $\overline{\overline{D}}$ le tenseur de dispersion (m^2.s^{-1}). Ce tenseur représente la superposition des phénomènes de diffusion moléculaire et de dispersion cinématique. On l'exprime par :

$$\overline{\overline{D_i}} = (\alpha_T \omega|u| + d_i) \cdot Id + (\alpha_L - \alpha_T)\omega \cdot \frac{\vec{u} \otimes \vec{u}}{|u|} \tag{1.77}$$

où α_L et α_T sont les dispersivités longitudinales et transversales du milieu (m) et d_i est le coefficient de diffusion moléculaire du soluté c_i.

D'autres possibilités existent ; ainsi Lichtner (1995) [114] utilise une équation de Nernst-Plank afin de tenir compte du champ électrique Φ ainsi que des différences spatiales de coefficients d'activité $grad(ln\gamma_i)$. Le flux s'exprime alors :

$$\vec{J}_i = \omega\vec{u}[c_i] - \tau\omega z_i \frac{d_i[c_i]}{RT} Fgrad(\Phi) - \tau\omega d_i (grad[c_i] + [c_i] grad(ln\gamma_i)) \tag{1.78}$$

où τ est la tortuosité du milieu, R la constante des gaz parfaits et F la constante de Faraday, T est la température, z_i est la charge électrique du soluté et γ_i est son coefficient d'activité.

1.3.3. TRANSPORT DE CHALEUR

1.3.3.1. Equation de transport de chaleur

Le transfert de chaleur en milieu poreux pose la question de la température de la phase fluide et de la phase solide. A priori, ces deux températures sont différentes et il est nécessaire de décrire à la fois les variations spatio-temporelles du champ de température de chacune des phases, mais également les échanges entre phases. Cette approche par non-équilibre local impose une description de l'interface liquide-solide (§ 1.1.1.6). Cette description a été étudiée par de nombreux auteurs [147, 150, 151, 166-170]. Un modèle simple inclus une équation comprenant des termes d'accumulation, d'advection, de diffusion et d'échanges entre phase solide et liquide [150] :

$$\begin{aligned}\frac{\partial}{\partial t}\left(\omega Cp_f \rho_f T_f\right) &= \nabla\left(\lambda_f \nabla T_f\right) - \nabla\left(\rho_f Cp_f \omega \cdot \vec{u} T_f\right) + a_h h\left(T_f - T_s\right) \\ \frac{\partial}{\partial t}\left[\left(1-\omega\right) Cp_s \rho_s T_s\right] &= \nabla\left(\lambda_s \nabla T_s\right) + a_h h\left(T_f - T_s\right)\end{aligned} \tag{1.79}$$

où T_f et T_s sont les températures de la phase fluide et solide respectivement ; et a_h et h sont la surface de contact entre les phases par unité de volume et h le coefficient de transfert thermique de l'interface fluide-solide. Cependant, ce modèle n'est pas dérivé d'une formulation à l'échelle du pore. Des modèles issus de formulation à l'échelle du pore existent [143, 150, 151, 170] mais sont plus complexes.

A l'échelle du VER, le transport de chaleur en milieu poreux est décrit par une équation d'advection-diffusion [77], similaire à celle décrivant le transport des solutés, si l'on suppose que l'équilibre thermique entre phase fluide et solide est rapidement atteint.

$$\frac{\partial}{\partial t}\left\{\left[\left(1-\omega\right) Cp_s \rho_s + \omega Cp_f \rho_f\right] T\right\} = \nabla\left(\lambda_{eq} \nabla T\right) - \nabla\left(\rho_f Cp_f \omega \cdot \vec{u} T\right) + \sum_{\substack{i \\ reactions}} \Delta H_i \cdot r_i \tag{1.80}$$

Le terme $\nabla\left(\lambda_{eq} \nabla T\right)$ représente la dispersion et la diffusion de la chaleur au sein du milieu poreux saturé, $\nabla\left(\rho_f Cp_f \omega \cdot \vec{u} T\right)$ représente le transport de chaleur par convection lié au mouvement de la phase fluide et $\sum_{\substack{i \\ reactions}} \Delta H_i \cdot r_i$ représente la production de chaleur due aux phénomènes chimiques. ΔH_i est l'enthalpie libre de la réaction i et r_i est la vitesse de la réaction.

Le tenseur de dispersion-diffusion de la chaleur, λ_{eq} se définit comme la somme d'un terme de dispersion, faisant intervenir le champ de vitesse de la phase fluide, et d'un terme de diffusion de la chaleur :

$$\lambda_{eq,i,j} = \omega \rho_f Cp_f \left[\alpha_T \left|\vec{u}\right| \delta_{ij} + \left(\alpha_L - \alpha_T\right) \frac{u_i u_j}{\left|\vec{u}\right|}\right] + \left[\left(1-\omega\right)\lambda_s + \omega \lambda_f\right] \delta_{ij} \tag{1.81}$$

On peut définir les grandeurs équivalentes suivantes pour le milieu poreux, capacité calorifique Cp_{PM} (1.82)et conductivité thermique λ_{PM} (1.83) :

$$Cp_{PM} = (1-\omega) Cp_s \rho_s + \omega Cp_f \rho_f \tag{1.82}$$

$$\lambda_{PM} = (1-\omega) \lambda_s + \omega \lambda_f \tag{1.83}$$

1.3.3.2. Approximation à porosité et densité constante

Si l'on suppose que les variations de porosité et de densité sont négligeables ou très lentes, on peut reformuler l'équation (1.80) pour obtenir une forme similaire à l'équation de transport de soluté. La capacité calorifique équivalente du milieu poreux (1.82) peut se reformuler en mettant en avant la partie *fluide* de ce terme :

$$Cp_{PM} = \omega Cp_f \rho_f \left(1 + \frac{(1-\omega) Cp_s \rho_s}{\omega Cp_f \rho_f} \right) = \omega Cp_f \rho_f \cdot R_T \tag{1.84}$$

où R_T apparait comme un facteur de retard thermique dû à la présence de la phase solide.

Ainsi, le terme d'accumulation de l'équation (1.80) s'écrit sous la forme suivante :

$$\frac{\partial}{\partial t} \left\{ \left[(1-\omega) Cp_s \rho_s + \omega Cp_f \rho_f \right] T \right\} = \omega Cp_f \rho_f \cdot R_T \frac{\partial T}{\partial t} \tag{1.85}$$

De même, le terme de convection thermique peut s'écrire dans ces conditions :

$$\nabla \left(\rho_f Cp_f \omega \cdot \vec{u} T \right) = \rho_f Cp_f \omega \cdot \nabla \left(\vec{u} T \right) \tag{1.86}$$

Le tenseur de dispersion-diffusion thermique se reformule de la façon suivante :

$$\lambda_{eq,i,j} = \omega \rho_f Cp_f \left\{ \left[\alpha_T \left| \vec{u} \right| \delta_{ij} + (\alpha_L - \alpha_T) \frac{u_i u_j}{\left| \vec{u} \right|} \right] + \frac{\left[(1-\omega) \lambda_s + \omega \lambda_w \right]}{\omega \rho_f Cp_f} \delta_{ij} \right\} = \omega \rho_f Cp_f \cdot \lambda_{eq,i,j} \tag{1.87}$$

L'équation de transfert de chaleur en milieu poreux prend alors la forme d'une équation d'advection-dispersion, similaire à celle du transport de soluté. Les outils numériques développés pour l'une peuvent être appliqué à l'autre très facilement :

$$R_T \frac{\partial T}{\partial t} = -\nabla \left(\vec{u} T \right) + \nabla \left(\lambda_{eq} \nabla T \right) + \frac{1}{\omega \rho_f Cp_f} \sum_{i}^{reactions} \Delta H_i \cdot r_i \tag{1.88}$$

1.3.4. CONCLUSION EQUATIONS DE CONSERVATION

Ainsi que cela a été abordé à plusieurs reprises, le point délicat concernant ces équations de conservation réside dans le changement d'échelle. Les équations écrites à l'échelle du pore comme celles écrites à l'échelle du VER sont bien connues, mais si les paramètres à l'échelle du pore sont relativement bien déterminés, le passage à l'échelle du VER introduit une grande variabilité : des termes complexes à mesurer comme la tortuosité ou la dispersivité apparaissent. Ces questions ont été largement abordées dans la littérature [150, 151, 170, 171], et le seront certainement encore longtemps.

Cependant, les travaux développés ici se placent résolument à l'échelle du VER et la problématique du niveau de description ne porte pas sur l'échelle spatiale mais plutôt sur l'aspect phénoménologique : quels sont les phénomènes à prendre en compte, quels sont ceux à négliger ? Parmi les points qui me semblent importants d'explorer, on trouve l'équation de Nernst-Planck (1.78) dans laquelle les termes électrostatiques en $grad(\Phi)$ et d'activité en $grad(ln\gamma)$ sont habituellement négligé. On retrouve alors l'équation d'advection-dispersion (1.76). Construire des critères objectifs permettant de déterminer dans quelles conditions l'un ou l'autre de ces termes peut être négligé constituera un axe de recherche.

1.4. REACTIONS

Afin de décrire les phénomènes chimiques, deux voies sont couramment utilisées : la description par cinétique chimique et la description par équilibre instantané. La description par cinétique chimique permet d'obtenir l'évolution d'un système chimique au cours du temps compte tenu des vitesses respectives des différentes réactions. La description par équilibre instantané donne la composition du système chimique après un temps suffisamment long pour qu'aucune modification de composition n'y soit plus détectable.

De ces deux définitions, il ressort que la description par équilibre est une approximation plus forte que celle par cinétique, et que tout système décrit de façon satisfaisante à l'aide d'une approche par équilibre pourra l'être également par cinétique.

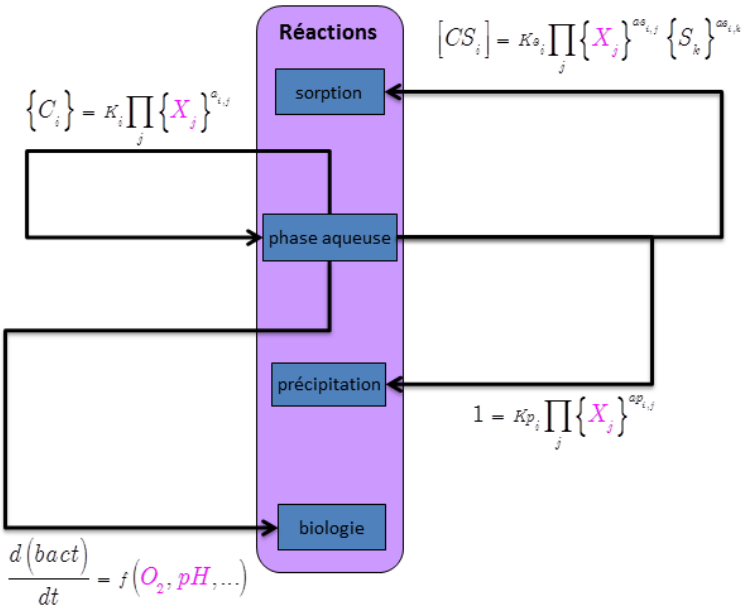


Figure 1-18 : Schéma des interactions liées aux réactions

1.4.1. CINETIQUE CHIMIQUE

Le choix de décrire les phénomènes chimiques par cinétique nécessite la connaissance des vitesses de réaction. Ces lois de vitesse peuvent prendre des formes extrêmement variées selon les phénomènes étudiés.

1.4.1.1. Vitesse de réaction

Dans l'équation de transport réactif (1.75), la vitesse des réactions, r_q, apparaît sous la forme de terme puit-source. Selon les phénomènes modélisés, l'écriture de ces vitesses de réaction peut prendre des formes très diverses.

Lorsque la réaction q est écrite sous la forme d'une réaction élémentaire :

$$\sum_i b_{q,i} c_i \underset{k_d}{\overset{k_f}{\rightleftharpoons}} \sum_j \tilde{b}_{q,i} \tilde{c}_i \tag{1.89}$$

où $\sum_i b_{q,i} c_i$ est la somme des réactifs multipliés par leur coefficient stœchiométrique et $\sum_j \tilde{b}_{q,i} \tilde{c}_i$ celle des produits de la réaction, et où k_f et k_d sont les constantes de vitesse de formation et de disparition des produits.

Dans ce cas, la vitesse de la réaction s'écrit :

$$r_q = k_f \prod_i [c_i]^{b_{q,i}} - k_d \prod_j \left[\tilde{c}_j\right]^{\tilde{b}_{q,j}} \tag{1.90}$$

Cette formulation est une forme fondamentale issue des mécanismes moléculaires en jeu lors de la réaction.

Cependant, les phénomènes chimiques modélisés actuellement en science de l'environnement sont souvent trop complexes pour être écrits sous la forme de réactions élémentaires. Des formulations spécifiques sont alors adoptées.

On peut, entre autre, présenter les formulations de type Monod, fréquemment utilisées pour décrire les phénomènes de croissance bactérienne. Ainsi, Wanko *et al.* [66] décrivent la croissance d'une biomasse en présence d'oxygène et d'un substrat carboné à l'aide de la loi suivante :

$$r = \left[k_1 \frac{[C_S]}{K_S + [C_S]} \cdot \frac{\left[O_{2,aq}\right]}{K_O + \left[O_{2,aq}\right]} - b \frac{\left[O_{2,aq}\right]}{K_O + \left[O_{2,aq}\right]} \right] k_2 \left[C_B\right] \tag{1.91}$$

où $\left[C_B\right]$ est la concentration en biomasse active (g.m^{-3}), $\left[O_{2,aq}\right]$ celle en oxygène dissout (g.m^{-3}) et $\left[C_S\right]$ celle du substrat carboné (g.m^{-3}). k_1 est la constante d'utilisation du substrat (s^{-1}), k_2 celle de l'utilisation de l'oxygène pour la synthèse de biomasse (-), b est la constante de respiration de la biomasse (s^{-1}). K_S est le coefficient de demi activité de Monod par rapport au substrat (g.m^{-3}) et K_O est le coefficient de demi activité de Monod par rapport à l'oxygène (g.m^{-3}).

Steefel et Lasaga (1994) décrivent la précipitation et dissolution des minéraux à l'aide de la loi suivante :

$$r = A \cdot k \cdot f(a) \cdot g(\Delta G) \tag{1.92}$$

où A est la surface réactive du minéral (m^2.m^{-3}), k est la constante de la réaction (mol.m^{-2}.s), $f(a)$ est une fonction des activités des ions en solution et $g(\Delta G)$ est une fonction de l'énergie libre de la solution.

Ces auteurs proposent la formulation suivante :

$$r = A \cdot k \cdot sgn\left(log \frac{Q_m}{K_m} \right) \cdot \prod_i \left(a_i^{\,p} \right) \cdot \left| \left(\frac{Q_m}{K_m} \right)^M - 1 \right|^N \tag{1.93}$$

avec $Q_m = \prod_i a_i^{b_{m,i}}$ l'indice de saturation du minéral ; K_m la constante de l'équilibre de précipitation-dissolution ; $\prod_i \left(a_i^{\,p} \right)$ le produit des activités des catalyseurs et inhibiteurs de la réaction, M et N sont des réels positifs empiriques.

On constate donc que les descriptions cinétiques des phénomènes chimiques sont très diverses et ne se plient pas à un formalisme unique. Elles constituent des systèmes couplés d'équations différentielles ordinaires (ODE) qui sont souvent qualifiés de raides, ou *stiff*. Ceci signifie que les temps caractéristiques des phénomènes, temps de demi-vie par exemple, varient sur plusieurs ordres de grandeur. Ainsi, une réaction acide-base en milieux aqueux est à l'équilibre en quelques nano - secondes, le doublement d'une population bactérienne nécessite quelques heures, alors que l'altération du quartz requière quelques milliers d'années.

1.4.1.1. Influence de la température

Pour les réactions inorganiques, il est classique de décrire l'influence de la température sur la vitesse de réaction par la loi d'Arrhénius :

$$\frac{d\left(ln\, k\right)}{dT} = \frac{E_a}{RT^2} \tag{1.94}$$

où k est la constante de vitesse d'une réaction et E_a son énergie d'activation (en J.mol^{-1}).

Cependant, dans le cas de réactions organiques, cette relation n'est plus valable, car les mécanismes en jeu ont un optimum thermique. A l'heure actuelle, nous n'avons pas trouvé de travaux compatibles avec la démarche de modélisation menée ici.

1.4.1.2. Bilan de matière

Dans le cas de phénomènes de transport réactif où la chimie est décrite par le biais de la cinétique, les vitesses de réactions apparaissent immédiatement dans l'équation bilan d'advection-dispersion-réaction (1.75).

1.4.2. CHIMIE A L'EQUILIBRE THERMODYNAMIQUE

1.4.2.1. Lois d'action de masse

1.4.2.1.1. Cas général

Considérons une réaction chimique, qui serait écrite sous forme de réaction élémentaire (1.89). La vitesse de cette réaction est donnée par la relation (1.90). Par définition, l'équilibre thermodynamique est atteint pour cette réaction lorsque sa vitesse est nulle :

$$r_q = k_f \prod_i [c_i]^{b_{q,i}} - k_d \prod_j \left[\tilde{c}_j\right]^{\tilde{b}_{q,j}} = 0 \tag{1.95}$$

Il n'y a alors plus d'évolution macroscopique de ce système. Il est habituel de réorganiser l'équation (1.95) sous la forme suivante qui fait apparaître la constante d'équilibre thermodynamique K_q :

$$\frac{\prod_j \left[\tilde{c}_j\right]^{\tilde{b}_{q,j}}}{\prod_i [c_i]^{b_{q,i}}} = \frac{k_f}{k_d} = K_q \tag{1.96}$$

1.4.2.1.2. Activité

Ecrite sous forme de loi d'action de masse, on obtient une formulation unique pour décrire l'ensemble des phénomènes à l'équilibre thermodynamique. Pour cela, on introduit la notion d'activité des espèces chimiques $\{c_i\}$. L'activité représente la capacité à réagir d'une espèce donnée, en tenant compte des autres espèces chimiques environnantes. Dans le cas idéal (gaz parfait, solution diluée) les activités sont données par le Tableau 1-7 :

Tableau 1-7 : Conventions d'activité

Condition	Activité
Solvant (eau)	1
Soluté dilué	Concentration molaire
Gaz	Pression partielle
Solide pur	1
Solution solide	Fraction molaire

Dans le cas réel, on peut être amené à introduire les coefficients d'activités γ_i présentés dans l'équation (1.78) afin de lier concentration ou pression partielle et activité :

$$\{c_i\} = \gamma_i [c_i] \quad \text{ou} \quad \{p_i\} = \gamma_i p_i \tag{1.97}$$

avec p_i la pression partielle du gaz.

En phase aqueuse, les coefficients d'activité se déterminent à partir de I qui est la force ionique de la solution :

$$I = \frac{1}{2}\sum_i z_i^{\,2}\left[c_i\right] \tag{1.98}$$

Lorsque la force ionique est inférieure à 0,5 ; ce qui est le cas de la majorité des eaux naturelles ; les coefficients d'activité peuvent se calculer à l'aide de la loi de Davies [172] :

$$log\left(\gamma_i\right) = -Az_i^{\,2}\left(\frac{\sqrt{I}}{1+\sqrt{I}} - BI\right) \tag{1.99}$$

Les paramètres de cette loi, A et B se déterminent de la façon suivante :

$$A = 1{,}82\cdot 10^6\left(\varepsilon_W T\right)^{-\frac{3}{2}} \text{ et } 0{,}2 \le B \le 0{,}3 \tag{1.100}$$

Depuis les travaux de Morel [173], la plupart des auteurs prennent $B = 0{,}24$.

La loi d'action de masse décrivant les équilibres chimiques s'écrit alors :

$$\frac{\prod_j \left\{\tilde{c}_j\right\}^{\tilde{a}_{q,j}}}{\prod_i \left\{c_i\right\}^{a_{q,i}}} = K_q \tag{1.101}$$

où les coefficients $a_{q,i}$ et $\tilde{a}_{q,i}$ ne sont pas forcément égaux aux coefficients stœchiométriques $b_{q,i}$ et $\tilde{b}_{q,i}$. Ce formalisme permet une description unique de la plupart des phénomènes chimiques.

Ainsi, pour la réaction en phase aqueuse de dissociation de l'acide carbonique :

$$H_2O + H_2CO_3 \rightleftarrows 2H_3O^+ + CO_3^{2-} \tag{1.102}$$

La loi d'action de masse s'écrit :

$$K = \frac{\left\{H_3O^+\right\}^2\cdot\left\{CO_3^{2-}\right\}}{\left\{H_2O\right\}\cdot\left\{H_2CO_3\right\}} \text{ ou } K = \frac{\gamma_1^{\,2}\left[H_3O^+\right]^2\cdot\gamma_2\left[CO_3^{2-}\right]}{\left[H_2CO_3\right]} \tag{1.103}$$

Les échanges liquide-gaz sont décrits par une loi de Henry, et pour la dissolution dans l'eau du dioxyde de carbone, on a :

$$CO_{2(gaz)} + H_2O_{(liq)} \rightleftarrows H_2CO_3 \tag{1.104}$$

La loi d'action de masse s'écrit :

$$K = \frac{\{H_2CO_3\}}{\{CO_{2(gaz)}\}\cdot\{H_2O_{(liq)}\}} = \frac{[H_2CO_3]}{P_{CO_{2(gaz)}}} \quad (1.105)$$

1.4.2.1.3. Echange d'ions

Les phénomènes d'échange d'ions peuvent être écrits sous un formalisme similaire. Un échange d'ion sodium calcium peut s'écrire :

$$2(\equiv S - Na) + Ca^{2+} \rightleftarrows (\equiv S_2 - Ca) + 2Na^{+} \quad (1.106)$$

$(\equiv S - Na)$ symbolise la surface du solide et met en évidence l'ion mobile, ici Na^{+}. La notation $(\equiv S_2 - Ca)$ indique clairement la stœchiométrie de la réaction et montre que l'ion divalent Ca^{2+} occupe deux sites de surface. La loi d'action de masse s'écrit :

$$K = \frac{[\equiv S_2 - Ca]\cdot\gamma_1^{\,2}\left[Na^{+}\right]^2}{\gamma_2\left[Ca^{2+}\right]\cdot[\equiv S - Na]^2} \quad (1.107)$$

1.4.2.1.4. Précipitation-dissolution

Les phénomènes de précipitation et dissolution s'insèrent également dans ce formalisme. Pour la précipitation de la calcite, on peut écrire :

$$Ca^{2+} + CO_3^{2-} \rightleftarrows CaCO_3 \quad (1.108)$$

Et la loi d'action de masse donne :

$$K = \frac{\gamma_2\left[Ca^{2+}\right]\cdot\gamma_2\left[CO_3^{2-}\right]}{\{CaCO_3\}_{solide}} = \gamma_2^{\,2}\left[Ca^{2+}\right]\cdot\left[CO_3^{2-}\right] \quad (1.109)$$

Dans l'écriture de la loi d'action de masse, l'activité de la calcite $\{CaCO_3\}$ est égale à 1 car il s'agit d'un solide pur. Cependant, la relation (1.109) n'est valable que lorsqu'il y a équilibre entre de la calcite solide et la phase aqueuse. Ainsi, il est préférable d'écrire cette loi d'action de masse sous la forme d'une inégalité :

$$K \geq \gamma_2^{\,2}\left[Ca^{2+}\right]\cdot\left[CO_3^{2-}\right] \quad (1.110)$$

1.4.2.1.5. Formation de solution solide

Lorsque plusieurs cations précipitent avec le même anion, il ne se forme pas plusieurs précipités distincts mais une solution solide. Alors, l'activité des espèces précipitées n'est plus égale à un mais doit être recalculée à partir de leurs fractions molaires respectives dans la solution solide formée. On peut ainsi avoir la formation de dolomite $Ca_xMg_{(1-x)}CO_3$ par précipitation simultanée de calcite $CaCO_3$ et de magnésite $MgCO_3$:

$$\begin{cases} CaCO_3 \rightleftarrows Ca^{2+} + CO_3^{2-} & K_1 \\ MgCO_3 \rightleftarrows Mg^{2+} + CO_3^{2-} & K_2 \end{cases} \quad (1.111)$$

La précipitation de la solution solide n'intervient que si la somme des produits de solubilité est supérieure ou égale à 1 :

$$\frac{1}{K_1}\left\{Ca^{2+}\right\}\left\{CO_3^{2-}\right\} + \frac{1}{K_2}\left\{Mg^{2+}\right\}\left\{CO_3^{2-}\right\} \geq 1 \quad (1.112)$$

On constate ainsi qu'il est possible d'inclure dans la solution solide un élément à l'état de trace en phase aqueuse. Lorsque la formation d'une solution solide survient, les lois d'action de masse décrivant la formation de chacun des précipités doit être modifiée pour prendre en compte le fait que le solide formé n'est pas pur, en faisant intervenir la fraction molaire des précipités.

$$\begin{cases} K_1 = \dfrac{\left\{Ca^{2+}\right\}\left\{CO_3^{2-}\right\}}{\dfrac{n_{CaCO_3}}{n_{CaCO_3} + n_{MgCO_3}}} \\ K_{-2} = \dfrac{\left\{Mg^{2+}\right\}\left\{CO_3^{2-}\right\}}{\dfrac{n_{MgCO_3}}{n_{CaCO_3} + n_{MgCO_3}}} \end{cases} \quad (1.113)$$

1.4.2.1.6. Complexation de surface

Les phénomènes de sorption par complexation de surface [173-175] prennent en compte la formation d'un potentiel électrique Ψ à proximité des surfaces lorsqu'une (ou des) espèce(s) chimique(s) est (sont) fixée(s). La réaction s'écrit, dans le cas de la protonation d'un oxyde de fer :

$$\left(\equiv Fe-OH\right) + H^+ \rightleftarrows \left(\equiv Fe-OH_2^+\right) \quad (1.114)$$

On comprend aisément que si la surface $\left(\equiv Fe-OH\right)$ est déjà en partie protonée, alors elle sera chargée positivement, et la fixation d'une espèce chargée positivement sera défavorisée. On montre alors [173, 176] que la loi d'action de masse relative cet équilibre fait intervenir le potentiel électrique de la surface :

$$K_{intr} \cdot exp\left(-\frac{\Delta z \cdot F}{RT}\Psi\right) = \frac{\left\{\equiv FeOH_2^+\right\}}{\left\{\equiv FeOH\right\}\left\{H^+\right\}} \quad (1.115)$$

où K_{intr} est la constante intrinsèque de l'équilibre, indépendante du potentiel électrique de surface ; F est la constante de Farraday, R la constante des gaz parfaits, T la température et Δz la différence de charge entre l'état de référence et l'état final de la surface.

Le potentiel de surface dépend de la charge électrique totale de la surface σ, qui se détermine par la relation suivante :

$$\sigma = \sum_{surface} z_i \left[c_i\right] \cdot \frac{F}{S \cdot \rho_S} \quad (1.116)$$

où l'on ne compte que les espèces c_i liées à la surface étudiée, S est la surface spécifique du matériau et ρ_S sa concentration massique dans le milieu poreux.

On utilise ensuite divers modèles pour exprimer le potentiel en fonction de la charge de surface [177]. Citons entre autre les modèles :

- à capacité constante (CCM)

$$\sigma = \frac{1}{cap} \cdot \Psi \tag{1.117}$$

- à couche diffuse (DLM)

$$\sigma = \left(8 \cdot RT \cdot \varepsilon \cdot \varepsilon_0 \cdot I\right)^{1/2} \cdot sinh\left(\frac{z_{el} \cdot F \cdot \Psi}{2 \cdot RT}\right) \tag{1.118}$$

De manière générale, les modèles de complexation de surface sont construits pour permettre de lier potentiel et charge de surface, ce que l'on peut symboliser par :

$$\sigma = f_{CS}\left(\Psi\right) \tag{1.119}$$

avec f_{CS} la fonction modèle de complexation de surface.

Bien que la théorie de la complexation de surface ne soit pas habituellement présentée sous cette forme, il est tout à fait possible d'interpréter l'influence du potentiel de surface comme le coefficient d'activité de la surface. La loi d'action de masse (1.115) s'écrit alors :

$$K_{intr} = \frac{\gamma_{surf}\left[FeOH_2^+\right]}{\left[\equiv FeOH\right]\gamma_1\left\{H^+\right\}} \tag{1.120}$$

Avec $\gamma_{surf} = exp\left(\frac{\Delta z \cdot F}{RT}\Psi\right)$ le coefficient d'activité de l'espèce chargée fixée sur la surface.

1.4.2.1.7. Précipitation de surface

Les phénomènes de précipitation de surface sont des phénomènes de sorption qui peuvent se produire lorsque la saturation des sites de surface n'est pas observée (Sigg *et al.* 2000 [176]). Plusieurs auteurs ont présenté ces phénomènes [177, 178]. A la surface d'un oxyde, $Ma(OH)_{3(s)}$, la précipitation de surface d'un cation métallique Me^{2+} conduit alors premièrement la formation d'un complexe de surface (1.121) puis la formation d'une solution solide.

$$\equiv\left\{Ma\left(OH\right)_{3(s)}\right\}-OH + Me^{2+} + 2H_2O \rightleftarrows Ma\left(OH\right)_{3(s)} + = \left\{Me\left(OH\right)_{2(s)}\right\}-OH_2^+ + H^+ \tag{1.121}$$

La précipitation de Me^{2+} s'écrit :

$$= \left\{Me\left(OH\right)_{2(s)}\right\}-OH_2^+ + Me^{2+} + 2H_2O \rightleftarrows Me\left(OH\right)_{2(s)} + = \left\{Me\left(OH\right)_{2(s)}\right\}-OH_2^+ + 2H^+ \tag{1.122}$$

La précipitation de Ma^{3+} s'écrit :

$$\equiv\left\{Ma\left(OH\right)_{3(s)}\right\}-OH+Ma^{3+}+3H_2O\rightleftarrows Ma\left(OH\right)_{3(s)}+\equiv\left\{Ma\left(OH\right)_{3(s)}\right\}-OH+3H^{+} \quad (1.123)$$

Les lois d'actions de masse relatives à la réaction de complexation de surface, (1.121), de précipitation de Me^{2+}, (1.122), et de précipitation de Ma^{3+}, (1.123), s'écrivent :

$$\left\{=\left\{Me\left(OH\right)_{2(s)}\right\}-OH_2^{+}\right\}=K_1\cdot e^{-\Psi}\cdot\frac{\left\{\equiv\left\{Ma\left(OH\right)_{3(s)}\right\}-OH\right\}\cdot\left\{Me^{2+}\right\}}{\left\{H^{+}\right\}} \quad (1.124)$$

$$\left\{Me\left(OH\right)_{2(s)}\right\}=K_2\cdot\frac{\left\{Me^{2+}\right\}}{\left\{H^{+}\right\}^2} \quad (1.125)$$

$$\left\{Ma\left(OH\right)_{3(s)}\right\}=K_3\cdot\frac{\left\{Ma^{2+}\right\}}{\left\{H^{+}\right\}^3} \quad (1.126)$$

1.4.2.1.8. Influence de la température et de la pression

Le formalisme des lois d'action de masse étant issu directement de l'application de la thermodynamique aux phénomènes chimiques, la prise en compte des modifications de température et de pression sur les équilibres chimiques est relativement aisé.

La loi de Van't Hoff permet d'exprimer la constante d'équilibre à la température T en fonction de l'enthalpie libre de la réaction $\Delta_r H^0$:

$$\frac{d}{dT}\left(ln\left[K\left(T\right)\right]\right)=\frac{\Delta_r H^0}{RT^2} \quad (1.127)$$

La pression a également une influence sur les équilibres chimiques. La relation suivante lie la constante d'équilibre aux variations de pression en faisant intervenir ΔV° qui est le changement de volume molaire partiel dû à la réaction :

$$\frac{d}{dp}\left(ln\,K\right)=-\frac{\Delta V^{\circ}}{RT} \quad (1.128)$$

1.4.2.2. Approche par composants

1.4.2.2.1. Définition

Depuis les travaux de Morel et Morgan [179], la plupart des modélisateurs [41, 61, 112, 114, 180] utilisent la représentation des systèmes chimiques par espèces et composants, proposée par ces

auteurs. L'utilisation du tableau des équilibres [179] permet alors une écriture claire et synthétique de l'ensemble des équilibres thermodynamiques du système chimique.

Parmi l'ensemble des *Nc* espèces chimiques c_i, présentes dans le système, on choisit *Nx* composants X_j de telle façon qu'il soit possible d'écrire la formation de toutes les espèces par une combinaison des composants et sans que l'on puisse former un composant par combinaison d'autres composants. Les composants X_j forment ainsi un système libre et générateur de l'ensemble des espèces c_i. Pour chaque espèce c_i, on écrit la réaction de formation de cette espèce à partir des composants choisis :

$$\sum_{j=1}^{Nx} b_{i,j} X_j \rightleftarrows c_i \quad \forall i = 1,...,Nc \tag{1.129}$$

où $b_{i,j}$ est le coefficient stœchiométrique pour la conservation de la matière du composant X_j lors de la formation de l'espèce c_i.

Le nombre de composants, *Nx*, servant à décrire un système chimique est donc égal à la différence entre le nombre d'espèces, *Nc*, et le nombre de réactions, *Nr*, soit à la variance *V* du système à température et pression fixées (Morel et Morgan, 1972 [179], Sigg *et al.*, 2000 [176]) :

$$Nx = Nc - Nr = V \tag{1.130}$$

De manière générale, les lois d'action de masse s'écrivent de la façon suivante :

$$\{c_i\} = K_i \prod_{j=1}^{Nx} \{X_j\}^{a_{i,j}} \tag{1.131}$$

où l'on exprime la formation de chacune des espèces à partir du jeu de composants choisis.

Le choix des *Nx* composants peut être assez libre, à la condition que le jeu de composants choisis fournisse un système libre et générateur de l'ensemble des espèces chimiques.

On définit alors $\left[T_j\right]$ la concentration totale en composant X_j ,

$$\left[T_j\right] = \left[X_j\right] + \sum_{i=1}^{Nc} b_{i,j} \cdot \left[c_i\right] \quad \forall j = 1,...Nx \tag{1.132}$$

L'équation (1.132) introduit une distinction entre la concentration en composant $\left[X_j\right]$ et les concentrations des autres espèces chimiques $\left[c_i\right]$. Or cette distinction n'est pas forcément nécessaire. On peut tout à fait considérer la réaction (1.133) pour chaque composant, qui a un coefficient stœchiométrique $b_{i,j} = 1$ et une constante d'équilibre $K_i = 1$:

$$X_j \rightleftarrows c_i \tag{1.133}$$

La concentration totale en composant s'écrit alors plus simplement :

$$\left[T_j\right] = \sum_{i=1}^{Nc} b_{i,j} \cdot \left[c_i\right] \quad \forall j = 1,...Nx \tag{1.134}$$

On fait également apparaitre une distinction entre les espèces mobiles $c_{i,mob}$ et les espèces immobiles $c_{i,im}$ en définissant pour chaque composant une concentration totale mobile $\left[T_{j,mob}\right]$ et une concentration totale immobile $\left[T_{j,im}\right]$, différente selon que le composant X_j est lui-même mobile $\left(X_{j,mob}\right)$ ou immobile $\left(X_{j,im}\right)$:

$$\left[T_{j,mob}\right] = \left[X_{j,mob}\right] + \sum_{i,mobile} b_{i,j} \cdot \left[c_{i,mob}\right] \text{ et } \left[T_{j,im}\right] = \sum_{i,immobile} b_{i,j} \cdot \left[c_{i,im}\right] \text{ si } X_j \text{ mobile} \tag{1.135}$$

$$\left[T_{j,mob}\right] = \sum_{i,mobile} b_{i,j} \cdot \left[c_{i,mob}\right] \text{ et } \left[T_{j,im}\right] = \left[X_{j,im}\right] \sum_{i,immobile} b_{i,j} \cdot \left[c_{i,im}\right] \text{ si } X_j \text{ immobile} \tag{1.136}$$

On a évidemment la relation suivante entre les différentes concentrations totales en composant :

$$\left[T_j\right] = \left[T_{j,mob}\right] + \left[T_{j,im}\right] \tag{1.137}$$

Il est possible de synthétiser ce formalisme chimique sous forme de tableau des équilibres [179] tel que présenté dans le Tableau 1-10. On constate que ce tableau a une structure par blocs, reflétant les interactions possibles ou non entre les différents composants. Ainsi, une surface réactive est décrite soit sous forme d'échangeur d'ions, soit sous forme de complexation de surface, et ne peut donc pas intervenir dans les réactions de l'autre catégorie. De même, on suppose ici que ces surfaces sont insolubles et donc qu'elles n'interviennent pas dans la formation d'espèces dissoutes ou gazeuses. Ici, les composants gazeux sont susceptibles d'être dissous dans la phase aqueuse et d'intervenir alors dans la formation de précipités. L'exemple le plus courant est le dioxyde de carbone, donnant par dissolution des carbonates qui vont précipiter avec du calcium ou du magnésium pour donner les roches carbonatées (calcaire, magnésie, dolomite).

1.4.2.2.2. Echange d'ions

La description des phénomènes d'échange d'ions dans le cadre d'un formalisme en composant ne présente pas de difficulté au premier abord, mais peut présenter quelques points délicats dans la pratique. Si l'on reprend l'exemple de la réaction (1.106), avec comme condition initiale une phase solide saturée en sodium à sa capacité d'échange cationique (CEC) en absence de calcium, on peut décrire ce système à l'aide du Tableau 1-8 :

Tableau 1-8 : Tableau des équilibres pour un échange d'ions avec un composant $\equiv S-Na$

	Ca^{2+}	Na^{+}	$\equiv S-Na$	Constante
Ca^{2+}	1	0	0	1
Na^{+}	0	1	0	1
$\equiv S-Na$	0	0	1	1
$\equiv S_2-Ca$	1	-2	1	K
Quantité totale	0	0	CEC	

Il est également possible de modifier le choix des composants pour choisir comme composant d'échange le solide saturé en calcium. La situation sera un peu plus compliquée, avec entre autre l'apparition de concentration totale négative, comme le montre le Tableau 1-9 :

Tableau 1-9 : Tableau des équilibres pour un échange d'ions avec un composant $\equiv S_2-Ca$

	Ca^{2+}	Na^{+}	$\equiv S_2-Ca$	Constante
Ca^{2+}	1	0	0	1
Na^{+}	0	1	0	1
$\equiv S-Na$	-1	2	1	1
$\equiv S_2-Ca$	0	0	1	$1/K$
Quantité totale	$-\frac{1}{2}CEC$	CEC	$\frac{1}{2}CEC$	

On constate sur les deux formulations présentées précédemment que l'une ou l'autre forme du solide est privilégiée. Le Tableau 1-8 est adapté à un solide qui reste majoritairement saturé en sodium alors que le Tableau 1-9 est adapté à un solide restant majoritairement saturé en calcium.

Tableau 1-10 : Tableau synthétique des équilibres

Espèce	Composants	gaz			aqueux			échange d'ions			Complexation de surface			Précipitation de surface			K
		Xg_1	...	Xg_{NxG}	Xa_1	...	Xa_{NxA}	Xe_1	...	Xe_{NxE}	Xs_1	...	Xs_{NxS}	Xsp_1	...	Xsp_{NxSP}	
gaz	cg_1	$a_{i,j}$; $b_{i,j}$			$a_{i,j}$; $b_{i,j}$			0			0			$a_{i,j}$; $b_{i,j}$			Constantes d'équilibre
	⋮																
	cg_{NcG}																
aqueux	ca_1	$a_{i,j}$; $b_{i,j}$			$a_{i,j}$; $b_{i,j}$			0			0			$a_{i,j}$; $b_{i,j}$			
	⋮																
	ca_{NcA}																
échange ion	ce_1	$a_{i,j}$; $b_{i,j}$			$a_{i,j}$; $b_{i,j}$			$a_{i,j}$; $b_{i,j}$			0			0			
	⋮																
	ce_{NcE}																
Complexation de surface	cs_1	$a_{i,j}$; $b_{i,j}$			$a_{i,j}$; $b_{i,j}$			0			$a_{i,j}$; $b_{i,j}$			0			
	⋮																
	cs_{NcS}																
Précipitation de surface	csp_1	$a_{i,j}$; $b_{i,j}$			$a_{i,j}$; $b_{i,j}$			0			0			$a_{i,j}$; $b_{i,j}$			
	⋮																
	csp_{NcSP}																
mineraux précipités	cp_1	$a_{i,j}$; $b_{i,j}$			$a_{i,j}$; $b_{i,j}$			0			0						
	⋮																
	cp_{NcP}																
Total initial		Tg_{init}			Ta_{init}			Te_{init}			Ts_{init}			Tsp_{init}			
X initial		Xg_{init}			Xa_{init}			Xe_{init}			Xs_{init}			Xsp_{init}			

On peut enfin choisir comme composant un solide fictif, n'existant pas réellement et donc n'apparaissant pas comme espèce, le solide nu $\equiv S-$. Le Tableau 1-11 montre alors que les 2 formes du solide, saturé en sodium ou saturé en calcium, sont considérées de façon identique.

Tableau 1-11 : Tableau des équilibres pour un échange d'ions avec un composant $\equiv S-$

	Ca^{2+}	Na^{+}	$\equiv S-$	Constante
Ca^{2+}	1	0	0	1
Na^{+}	0	1	0	1
$\equiv S-Na$	0	1	1	K_{Na}
$\equiv S_2-Ca$	1	0	2	K_{Ca}
Quantité totale	$-\frac{1}{2}CEC$	CEC	$\frac{1}{2}CEC$	

On a alors la relation suivante entre les deux constantes d'équilibre, dont l'une devra être fixée arbitrairement :

$$K = \frac{K_{Ca}}{K_{Na}^{\ 2}} \tag{1.138}$$

1.4.2.2.3. Précipitation - dissolution

Les réactions de précipitation-dissolution imposent une modification de la forme des lois d'action de masse. En effet, pour les autres réactions, la concentration des espèces est donnée par la loi d'action de masse (aux corrections d'activité près). Dans le cas des réactions de précipitation-dissolution, la loi d'action de masse donne l'activité du solide, qui ne peut être reliée à une quantité de matière :

$$1 \geq K_i \prod_{j=1}^{Nx} \left\{ X_j \right\}^{a_{i,j}} \tag{1.139}$$

1.4.2.2.4. Formation de solution solide

Afin de décrire la formation d'une solution solide (eq. (1.113)), il est nécessaire d'ajouter les nombres de moles de chacun des précipité impliqués (n_{Pi}) dans la solution solide au jeu des composants :

$$1 \geq K_i \frac{\prod_j \left\{ X_j \right\}^{a_{j,i}}}{\frac{n_{Pi}}{\sum_k n_{Pk}}} \tag{1.140}$$

1.4.2.2.5. Complexation de surface.

Si l'on considère une réaction de complexation de surface écrite de façon générique entre une surface $\equiv S$ et un complexant M :

$$\equiv S-OH + M^{2+} \rightleftarrows \equiv S-OM^{+} + H^{+} \quad K = \frac{\left\{\equiv S-OM^{+}\right\}\cdot\left\{H^{+}\right\}}{\left\{\equiv S-OH\right\}\cdot\left\{M^{2+}\right\}} \tag{1.141}$$

De façon plus générique, une telle réaction s'écrit :

$$cs_i = \sum_j b_{i,j} X_j \tag{1.142}$$

En construisant un composant électrostatique fictif, noté X_Ψ ,

$$\left\{X_\Psi\right\} = exp\left(-\frac{F}{RT^\circ}\cdot\Psi\right) \tag{1.143}$$

on peut alors faire intervenir la partie électrostatique de la constante d'équilibre telle que présentée dans l'équation (1.115) dans une loi d'action de masse exprimée en composant :

$$\left\{\equiv S-OM^{+}\right\} = K_{intr}\left\{X_\Psi\right\}\frac{\left\{\equiv S-OH\right\}\cdot\left\{M^{2+}\right\}}{\left\{H^{+}\right\}} \tag{1.144}$$

Ou de façon plus générique :

$$\left\{cs_i\right\} = K_{intr,i}\left\{X_\Psi\right\}\prod_j \left\{X_j\right\}^{a_{i,j}} \tag{1.145}$$

L'ajout de ce nouveau composant impose la prise en compte d'une nouvelle équation de conservation. Celle-ci sera utilisée pour déterminer la quantité totale de charge sur la surface (1.146) en ne prenant en compte que les espèces fixées sur la surface $\equiv S$.

$$\left[T_\Psi\right] = \sum_{\substack{i \\ surface}} z_i\left[c_i\right] = \sum_i b_{i,\Psi}\left[c_i\right] \tag{1.146}$$

Pour peu que l'on impose $b_{i,\Psi} = z_i$ si l'espèce est sur $\equiv S$ et $b_{i,\Psi} = 0$ sinon, l'équation de conservation (1.146) ne diffère pas des autres équations de conservation.

La relation (1.116) permet de calculer la charge électrique de surface en fonction de la quantité totale de charge sur la surface. Ensuite, à l'aide d'un des modèle ; CCM décrit par l'équation (1.117) ou DLM décrit par (1.118) ; on peut déterminer le potentiel de surface Ψ. En utilisant le formalisme générique fourni par la relation (1.119), cela se traduit par une relation entre le potentiel et la quantité de charge sur la surface :

$$\left[T_\Psi\right] = \frac{S\cdot\rho_S}{F}\cdot f_{CS}\left(\Psi\right) \text{ ou bien } \left[T_\Psi\right] = \frac{S\cdot\rho_S}{F}\cdot f_{CS}\left(-\frac{RT^\circ}{F} ln\left\{X_\Psi\right\}\right) \tag{1.147}$$

1.4.2.2.6. Précipitation de surface

Comme cela a été présenté (§ 1.4.2.1.5), la précipitation de $Ma(OH)_{3(s)}$ et de $Me(OH)_{2(s)}$ forme une seule phase solide. L'activité de cette phase solide est égale à 1, mais l'activité de chaque précipité est égale à sa fraction molaire dans la phase solide.

$$\left\{Ma(OH)_{3(s)}\right\} = \frac{\left[Ma(OH)_{3(s)}\right]}{\left[Ma(OH)_{3(s)}\right] + \left[Me(OH)_{2(s)}\right]} \tag{1.148}$$

$$\left\{Me(OH)_{2(s)}\right\} = \frac{\left[Me(OH)_{2(s)}\right]}{\left[Ma(OH)_{3(s)}\right] + \left[Me(OH)_{2(s)}\right]} \tag{1.149}$$

Un composant de précipitation de surface, T_P, est défini de la façon suivante :

$$[T_P] = \left[Ma(OH)_{3(s)}\right] + \left[Me(OH)_{2(s)}\right] \tag{1.150}$$

Les lois d'action de masse (1.125) et (1.126) sont réécrites en utilisant le composant de précipitation de surface.

$$\left[Me(OH)_{2(s)}\right] = K_2 \cdot \frac{\left\{Me^{2+}\right\}}{\left\{H^+\right\}^2} \cdot [T_P] \tag{1.151}$$

$$\left[Ma(OH)_{3(s)}\right] = K_3 \cdot \frac{\left\{Ma^{2+}\right\}}{\left\{H^+\right\}^3} \cdot [T_P] \tag{1.152}$$

Afin de respecter l'activité de la phase solide égale à 1, la loi de conservation pour T_P est écrite :

$$[T_P] = \left\{Ma(OH)_{3(s)}\right\} + \left\{Me(OH)_{2(s)}\right\} = 1 \tag{1.153}$$

La relation (1.153) impose de distinguer en temps qu'espèces chimiques l'activité et la concentration des précipités. Le système chimique décrit par les réactions (1.121), (1.122) et (1.123) est représenté par le Tableau 1-12. Le Tableau 1-12 fait intervenir les deux matrices stœchiométriques **B** pour les lois de conservation et **A** pour les lois d'action de masse. Les coefficients stœchiométriques **B** des espèces représentées par les activités des précipités sont nuls. Ainsi, ces espèces n'interviennent pas dans le bilan de matière : ce sont des espèces virtuelles.

Tableau 1-12 : Tableau des équilibres pour précipitation de surface
Lorsque 2 coefficients stœchiométriques sont donnés, lire $a_{i,j}$, $b_{i,j}$

	H^+	Ma^{3+}	Me^{2+}	$\equiv\left\{Ma\left(OH\right)_{3(s)}\right\}-OH$	X_ψ	T_P	log (K)
H^+	1						0
Ma^{3+}		1					0
Me^{2+}			1				0
$\equiv\left\{Ma\left(OH\right)_{3(s)}\right\}-OH$				1			0
$=\left\{Me\left(OH\right)_{2(s)}\right\}-OH_2^+$	-1		1	1	1		log (K_1)
[$Ma(OH)_{3(s)}$]	-3	1				1 , 0	log (K_3)
[$Me(OH)_{2(s)}$]	-2		1			1 , 0	log (K_2)
{$Ma(OH)_{3(s)}$}	-3 , 0	1 , 0				0 , 1	log (K_3)
{$Me(OH)_{2(s)}$}	-2 , 0		1 , 0			0 , 1	log (K_2)
	[T_H]	[T_{Ma}]	[T_{Me}]	[T_{Surf}]		1	

1.4.3. FORMULATION DU TRANSPORT REACTIF

1.4.3.1. Hypothèse de l'équilibre instantané

Lorsque l'on souhaite étudier des phénomènes chimiques en faisant l'hypothèse de l'équilibre instantané, Il est nécessaire de reformuler le bilan de matière présenté par l'équation d'advection-dispersion-réaction (1.75). En effet, dans ce bilan de matière, les réactions chimiques apparaissent sous forme de quantité de matière consommé par unité de temps ($-\sum_{réactions,r} b_{i,r}\cdot r_r$). Or le principe d'une description par équilibre instantané est d'éliminer la variable temporelle en supposant que les réactions chimiques sont infiniment rapides. La méthode utilisée [114, 181] pour modéliser les phénomènes de transport réactif à l'équilibre instantané est la suivante :

Pour les espèces, on peut réécrire l'équation d'advection-dispersion-réaction (ADR) en modifiant le terme de réaction, car dans la formulation par composant, une espèce chimique n'intervient que dans une seule réaction : celle qui décrit sa propre formation à partir du jeu de composants choisi. On fait également apparaître une distinction selon que l'espèce est mobile ou non :

$$\frac{\partial}{\partial t}\left(\omega\left[c_{i,mob}\right]\right)+div\left(\vec{J}_i\right)=-r_i \text{ si } c_i \text{ mobile} \tag{1.154}$$

$$\frac{\partial}{\partial t}\left(\omega\left[c_{i,im}\right]\right)=-r_i \text{ si } c_i \text{ immobile} \tag{1.155}$$

Pour les composants, le terme de réaction r_j reprend la somme de tous les termes réactionnels des espèces r_i multipliés par leurs coefficients stœchiométriques $b_{i,j}$, car les composants interviennent comme réactifs dans les réactions de formation des espèces. On a alors :

$$r_j = -\sum_{i=1}^{Nc} b_{i,j} \cdot r_i \quad \forall j = 1,...Nx \tag{1.156}$$

L'équation d'advection-dispersion réaction devient :

$$\frac{\partial}{\partial t}\left(\omega\left[X_{j,mob}\right]\right) + div\left(\vec{J}_j\right) = \sum_{i=1}^{Nc} b_{i,j} \cdot r_i \text{ si } X_j \text{ mobile} \tag{1.157}$$

$$\frac{\partial}{\partial t}\left(\omega\left[X_{j,im}\right]\right) = \sum_{i=1}^{Nc} b_{i,j} \cdot r_i \text{ si } X_j \text{ immobile} \tag{1.158}$$

Si l'on somme l'équation d'ADR pour un composant avec les équations d'ADR pour toutes les espèces multipliées par leurs coefficients stœchiométriques respectifs, on a alors :

$$\begin{aligned} &\frac{\partial}{\partial t}\left(\omega\left[X_{j,mob}\right]\right) + \sum_{i,mobile}\left\{b_{i,j} \cdot \frac{\partial}{\partial t}\left(\omega\left[c_{i,mob}\right]\right)\right\} + \sum_{i,immobile}\left\{b_{i,j} \cdot \frac{\partial}{\partial t}\left(\omega\left[c_{i,im}\right]\right)\right\} + div\left(\vec{J}_j\right) + \sum_{i,mobile}\left\{b_{i,j} \cdot div\left(\vec{J}_i\right)\right\} \\ &= \sum_{i=1}^{Nc} b_{i,j} \cdot r_i - \sum_{i,mobile}\left\{b_{i,j} \cdot r_i\right\} - \sum_{i,immobile}\left\{b_{i,j} \cdot r_i\right\} \end{aligned} \tag{1.159}$$

Ce qui donne, en exploitant la linéarité des opérateurs de dérivée spatiale et temporelle :

$$\frac{\partial}{\partial t}\left(\omega\left[X_{j,mob}\right] + \omega \sum_{i,mobile} b_{i,j} \cdot \left[c_{i,mob}\right] + \omega \sum_{i,immobile} b_{i,j} \cdot \left[c_{i,im}\right]\right) + div\left(\vec{J}_j + \sum_{i,mobile} b_{i,j} \cdot \vec{J}_i\right) = 0 \tag{1.160}$$

Cette nouvelle forme de l'équation d'ADR fait apparaître les concentrations totales en composant et élimine les termes réactionnels.

$$\frac{\partial}{\partial t}\left(\omega\left[T_j\right]\right) + div\left(\vec{J}_j + \sum_{i,mobile} b_{i,j} \cdot \vec{J}_i\right) = 0 \text{ si } X_j \text{ mobile} \tag{1.161}$$

Lorsque le composant X_j est immobile, l'équation devient :

$$\frac{\partial}{\partial t}\left(\omega\left[T_j\right]\right) + div\left(\sum_{i,mobile} b_{i,j} \cdot \vec{J}_i\right) = 0 \text{ si } X_j \text{ immobile} \tag{1.162}$$

Par souci de simplification, on écrit parfois l'équation d'ADR sous la forme suivante :

$$\frac{\partial}{\partial t}\left(\omega\left[T_j\right]\right) + div\left(J_{tot,j}\right) = 0 \tag{1.163}$$

1.4.3.2. Formulation mixte cinétique-équilibre

Dans certaines conditions, il peut être intéressant de décrire un système chimique en utilisant simultanément les deux formulations, cinétique et équilibre thermodynamique. Cette formulation mixte a été utilisée par différents auteurs pour décrire des phénomènes de diagénèse [40, 131, 137, 182] ou bien des phénomènes biologiques [67, 68, 183, 184] Ce type de système était proposé pour le niveau le plus difficile du benchmark MoMaS [98]. Nous présentons ici la méthodologie développée pour associer ces deux représentations en une formulation cohérente.

Pour obtenir un système d'équations couplant transport, chimie à l'équilibre et chimie cinétique, il est nécessaire de considérer les équations (1.154) et (1.155) qui constituent la formulation élémentaire du transport de chaque espèce. On utilisera également la décomposition du système chimique à l'équilibre entre espèces et composants ainsi que les équation de conservation de chaque composant (1.157) et (1.158). Il faut néanmoins introduire la distinction entre les espèces dont la formation sera décrite par équilibre et celles dont la formation sera décrite par cinétique. Il est évident que les composants doivent être choisis parmi les espèces décrites à l'équilibre.

Les réactions chimiques modélisées peuvent donc s'écrire sous la forme suivante pour les réactions à l'équilibre :

$$c_i \rightleftarrows \sum_j b_{i,j} X_j \tag{1.129}$$

et pour les réaction décrites par cinétique, en notant cc_n les espèces décrites par cinétique :

$$cc_n \rightleftharpoons \sum_k bc_{n,k;cin} \cdot cc_k + \sum_i bc_{n,i;eq} \cdot c_n + \sum_m bx_{n,m;eq} \cdot X_m \tag{1.164}$$

Il faut alors modifier les équations de conservation des espèces qui seront décrites par équilibre de la manière suivante :

$$\frac{\partial}{\partial t}\left(\omega\left[c_{i;mob}\right]\right) + div\left(\vec{J}_i\right) = -r_{i;eq} - r_{i;cin} \text{ si } c_i \text{ mobile} \tag{1.165}$$

$$\frac{\partial}{\partial t}\left(\omega\left[c_{i;im}\right]\right) = -r_{i;eq} - r_{i;cin} \text{ si } c_i \text{ immobile} \tag{1.166}$$

où $r_{i;eq}$ est la vitesse des réactions qui seront décrites par équilibre et $r_{i;cin}$ est celle des réactions qui seront décrites par cinétique.

Pour les espèces cinétiques, elles ne seront impliquées que dans des réactions décrites par cinétique :

$$\frac{\partial}{\partial t}\left(\omega\left[cc_{n;mob}\right]\right) + div\left(\vec{J}_n\right) = -rc_{n;cin} \text{ si } cc_n \text{ mobile, équilibre} \tag{1.167}$$

$$\frac{\partial}{\partial t}\left(\omega\left[cc_{n;im}\right]\right) = -rc_{n;cin} \text{ si } cc_i \text{ immobile, équilibre} \tag{1.168}$$

Le même formalisme est applicable aux composants choisis, ce qui donne :

$$\frac{\partial}{\partial t}\left(\omega\left[X_{j;mob}\right]\right) + div\left(\vec{J}_j\right) = -r_{j;eq} - r_{j;cin} \text{ si } X_j \text{ mobile} \tag{1.169}$$

$$\frac{\partial}{\partial t}\left(\omega\left[X_{j;im}\right]\right) = -r_{j;eq} - r_{j;cin} \text{ si } X_j \text{ immobile} \tag{1.170}$$

On a alors les relations suivantes entre les différentes vitesses des réactions ; pour les espèces et composants à l'équilibre, la relation (1.156) reste valable ;

$$r_{j;eq} = -\sum_i b_{i,j} \cdot r_{i;eq} \tag{1.156}$$

Pour les espèces à l'équilibre et les espèces cinétiques, on a :

$$r_{i;cin} = -\sum_n bc_{n,i;eq} \cdot rc_n \tag{1.171}$$

Pour les composants à l'équilibre et les espèces cinétiques, on a :

$$r_{j;cin} = -\sum_n bx_{n,j;eq} \cdot rc_n \tag{1.172}$$

En sommant la relation (1.169) (resp. (1.170) avec les équations de conservation des espèces décrites par équilibre multipliées par leurs coefficients stœchiométriques respectifs (1.165) (resp. (1.166), on obtient les équations de conservation des totaux en composants mobiles et immobiles :

$$\frac{\partial}{\partial t}\left(\omega\left[T_j\right]\right) + div\left(\vec{J}_j + \sum_{i;mobile} b_{i,j} \cdot \vec{J}_i\right) = \sum_n bx_{n,j;eq} \cdot rc_n + \sum_i b_{i,j} \cdot \left(\sum_n bc_{n,i;eq} \cdot rc_n\right) \text{ si } X_j \text{ mobile} \tag{1.173}$$

$$\frac{\partial}{\partial t}\left(\omega\left[T_j\right]\right) + div\left(\sum_{i,mobile} b_{i,j} \cdot \vec{J}_i\right) = \sum_n bx_{n,j;eq} \cdot rc_n + \sum_i b_{i,j} \cdot \left(\sum_n bc_{n,i;eq} \cdot rc_n\right) \text{ si } X_j \text{ immobile} \tag{1.174}$$

Les équations de conservations des espèces cinétiques mobiles (1.167) et immobiles (1.168) ainsi que celles des composants mobiles (1.173) et immobiles (1.174) sont similaires. Elles dépendent toutes du flux de matière $\vec{J}$ de chacune des grandeurs conservées et des vitesses des réactions que l'on choisit de décrire par cinétique rc_n.

1.4.4. CONCLUSION REACTIONS

Nous avons montré que la description des phénomènes chimiques dans les solutions aqueuses diluées était un domaine arrivé à maturité. Les deux formalismes, à l'équilibre instantané et sous forme cinétique sont à même de cohabiter au sein de la représentation d'un système chimique complexe. Ces formalismes sont également compatibles avec les équations de transport permettant ainsi la construction de modèles de transport réactifs fondés sur des lois phénoménologiques.

Cependant, nous avons restreint cette présentation aux solutions diluées, dont la force ionique est inférieure à 0,5 mol/L^{-1}. Certains codes de transport réactifs [185, 186] ont intégré le modèle de Pitzer pour calculer les corrections d'activité offrant la possibilité de décrire des solutions plus concentrées.

2. DEVELOPPEMENTS NUMERIQUES

Un modèle de transport réactif doit relever plusieurs défis, souvent fortement contradictoires :

i) Le défi de la précision. Il s'agit de résoudre le système d'équations décrivant les phénomènes avec le minimum d'erreurs. L'une des principales difficultés pour évaluer la performance d'un code au regard de la précision vient de l'absence de solution analytique dès que le problème de transport réactif devient un tant soit peu complexe.

ii) Le défi de la robustesse. Il arrive fréquemment qu'au cours d'un processus de résolution, le code se bloque, par division par zéro (ou nombre trop petit), par multiplication par l'infini (ou nombre trop grand), par non convergence d'un processus itératif… Un code est dit robuste lorsqu'il surmonte ce type de difficultés.

iii) Le défi de la rapidité. Les calculs de transport réactif sont souvent longs. Il est important de pouvoir effectuer ces calculs rapidement.

iv) Le défi de la flexibilité. Le développement d'un code de transport réactif est un travail long et complexe. Il est donc intéressant de concevoir un code qui puisse évoluer et intégrer de nouveaux phénomènes, utiliser de nouvelles méthodes numériques, moyennant un minimum de modifications.

2.1. ECOULEMENT ET TRANSPORT

2.1.1. CONTEXTE

Historiquement, les travaux, développés au sein de l'Institut de Mécanique des Fluides de Strasbourg puis du Laboratoire d'Hydrogéologie et de Géochimie de Strasbourg sur les modèles d'écoulement et de transport en milieux poreux par Ph. Ackerer, R. Mosé, A. Younès, H. Hoteit, F. Lehmann, ont permis l'émergence d'un modèle d'écoulement et de transport en milieu poreux : TRACES [13, 187-189] basé sur une combinaison d'éléments finis mixtes hybrides et d'éléments finis discontinus. Les éléments finis discontinus sont utilisés pour résoudre la partie advective de l'équation de transport et les éléments finis mixtes hybrides pour résoudre l'équation d'écoulement ainsi que la partie dispersive de l'équation de transport.

Les éléments finis mixtes hybrides permettent une conservation rigoureuse du bilan de masse sur chaque élément, la prise en compte d'un tenseur de perméabilité ou de dispersivité plein et garantissent un schéma inconditionnellement stable.

Les éléments finis discontinus permettent de modéliser des fronts de concentration très raides et garantissent l'absence d'oscillations. Cependant, leur formulation explicite en temps nécessite un respect strict du critère de Courant Friedrich Lévy (CFL).

2.1.2. SCHEMA ITERATIF ASSOCIE A UNE COMBINAISON D'ELEMENTS FINIS DISCONTINUS ET MIXTES HYBRIDES

L'intérêt de cette approche se retrouve tout au long des travaux réalisés, comme par exemple dans la précision et l'efficacité du code présenté lors du benchmark Transport Réactif MoMaS [84, 99]. Le travail présenté en Annexe 1 ([87]) illustre bien l'intérêt de cette combinaison *éléments finis dicontinus – éléments finis mixtes hybrides* dans le contexte du transport réactif. L'expérience modélisée, réalisée par Lefèvre *et al.* [190], consiste en l'injection de Strontium dans une colonne remplie de sable calcaire. Les réactions de précipitation, d'échange d'ions et de dissolution qui se produisent au sein du milieu entrainent la formation d'un front compressif se traduisant par un pic de strontium en sortie de colonne. Ce pic est, du fait des réactions, retardé, plus bref dans le temps et d'amplitude plus importante que celui observé pour un soluté non réactif.

Nous montrons que le schéma numérique utilisé permet la modélisation de ce front compressif avec une discrétisation spatiale et temporelle relativement modérée. Sur cet exemple, il est en effet impératif de maitriser au mieux la diffusion numérique, car celle-ci a un impact prépondérant sur la solution. Les phénomènes de précipitations, à l'origine de l'amplification du pic de Strontium, sont des phénomènes *à seuil* : ils ne se produisent que si la concentration dépasse le seuil de saturation et conduisent alors à une accumulation locale de strontium dans une zone de la colonne. Une diffusion numérique trop importante conduit à une baisse des concentrations, qui peuvent alors être inférieures au seuil de saturation. Il n'y a alors pas de précipitation et donc pas d'amplification du pic de strontium.

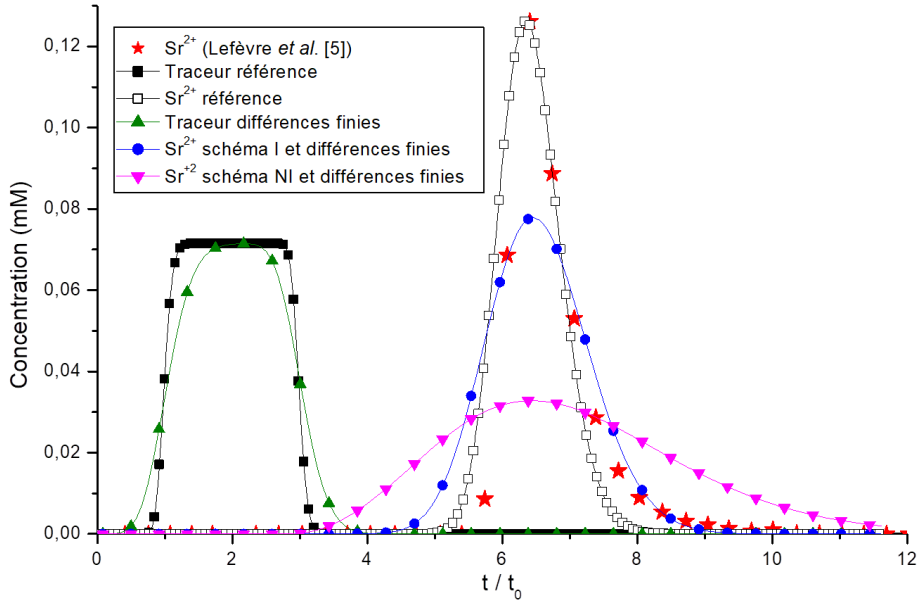


Figure 2-1 : Courbes d'élution du Strontium expérimentales et modélisées par différences finies.
(expérience d'après Lefèvre et al. [190])

Ainsi, bien que ces résultats ne soient pas présentés dans l'article [87], nous pouvons observer dans la Figure 2-1 qu'un modèle par différence finie est extrêmement sensible au choix de l'algorithme de couplage et que l'amplification du pic de strontium est complètement perdue si le couplage est réalisé par un schéma NI. Au contraire, on peut constater (voir Annexe 1) qu'un modèle basé sur des éléments finis discontinus et mixtes hybrides conduit à des résultats relativement proches, que le schéma de couplage soit itératif ou non, et que l'amplification du pic de strontium est correctement rendue.

2.2. COUPLAGE TRANSPORT – CHIMIE

2.2.1. CONTEXTE

Depuis les travaux de Yeh et Tripathi [83], les codes de transport réactif sont classés en 3 catégories, selon la méthode utilisée pour associer transport et chimie.

(i) L'approche globale consiste à écrire un ensemble d'équations décrivant à la fois le transport des différents solutés présents et les réactions chimiques ayant lieux dans le domaine. Yeh et Tripathi avaient montré que cette approche était grande consommatrice de ressources informatiques, à la fois en temps de calcul et en place mémoire. Actuellement, la réduction des temps de calcul demeure un enjeu important pour la modélisation du transport réactif. Mais grâce à l'évolution des technologies et la baisse des coûts de la mémoire informatique il n'y a plus réellement de pression pour réduire la consommation en place mémoire. De plus, de nombreux travaux récents mettent en évidence de nouvelles méthodes numériques conduisant à des codes de transport réactif par approche globale aussi rapide, sinon plus, que les autres.

(ii) L'approche par séparation d'opérateur non itérative consiste en une résolution successive des opérateurs de transport puis de chimie. Ainsi, au cours d'un pas de temps, on peut tout d'abord résoudre l'opérateur de transport pour chacune des espèces chimiques, puis résoudre l'opérateur de chimie dans chacune des mailles du domaine. Cette approche est extrêmement rapide, et peu gourmande en place mémoire. Cependant, la séparation des opérateurs introduit des erreurs intrinsèques, appelées erreurs de séparation d'opérateurs.

(iii) L'approche par séparation d'opérateur itérative conserve la philosophie générale de la séparation des opérateurs, mais on rajoute des termes correctifs dans chacun des opérateurs pour annuler, ou minimiser les erreurs de séparation d'opérateurs. Ces termes correctifs sont calculés par itérations entre transport et chimie.

D'un point de vue général, la résolution d'un problème de transport réactif s'apparente à la résolution d'un système non linéaire. Les approches classiques en mathématiques pour résoudre de tels systèmes sont des méthodes d'ordre zéro (type point fixe) et des méthodes d'ordre un (type Newton). On peut ainsi apparenter une approche globale à une méthode de Newton et une approche par séparation d'opérateurs itérative à une méthode du point fixe. L'approche par séparation d'opérateurs non itérative peut s'apparenter à la première itération du point fixe.

En termes de mise en œuvre opérationnelle (développement, mise à jour, amélioration) du code, l'approche globale est assez lourde alors que les approches par séparation d'opérateur, plus modulaires, sont beaucoup plus souples. Compte tenu de la complexité et de la multiplicité des phénomènes envisagés dans ce code, nous avons opté pour une approche par séparation d'opérateurs.

2.2.2. SEPARATION D'OPERATEURS ET ERREURS EN BILAN DE MASSE

Il est désormais bien établi que la résolution par séparation d'opérateurs peut introduire des erreurs intrinsèques [88, 191-199], liées à la résolution séparée d'opérateurs mathématiques représentant des phénomènes physico-chimiques fondamentalement liés. Dans le travail présenté en Annexe 2, nous présentons une compilation des différents schémas de séparation d'opérateurs recensés, itératifs et non itératifs, et nous établissons une étude analytique des erreurs de séparation d'opérateurs. En effet, lors d'une étude numérique, il est assez complexe de faire la part entre les erreurs de troncature (en temps et en espace), les erreurs de convergence (lors d'une résolution itérative de la chimie) et les erreurs de séparation d'opérateurs. Une étude analytique offre l'avantage d'éliminer erreurs de troncature et de convergence. Cependant, une telle étude n'est possible que sur un problème simplifié.

Il s'agit donc à un système chimique composé de 2 espèces interagissant par une réaction réversible décrite par une cinétique d'ordre 1. Le domaine étudié est mono-directionnel, semi infini, homogène, avec une condition de flux imposé à la frontière amont. Dans ces conditions, le bilan de matière pour chaque élément chimique peut être calculé analytiquement, pour une résolution globale comme pour une résolution par séparation d'opérateur ; et ce à chaque pas de temps et à chaque itération (pour les schémas de séparation itératifs).

Les erreurs de séparation d'opérateurs ainsi que les vitesses de convergence des schémas itératifs dépendent alors de temps adimensionnels construit à partir des constantes de réaction et des pas de temps des schémas de séparation. Certains schémas de séparation se révèlent alors plus imprécis (erreur de séparation forte) et/ou inefficaces (vitesse de convergence lente) que d'autres. Pour les schémas non itératifs, le schéma Strang-Splitting conduit aux plus faibles erreurs de bilan de masse ; pour les schémas itératifs, le schéma Symmetric est à préférer car il n'introduit pas d'erreurs de séparation et converge en 2 itérations (voir Annexe 2).

Cependant, ces résultats obtenus dans le cadre d'une étude analytique doivent être modérés et adaptés aux situations réelles. Ainsi, le schéma Strang-splitting conduit bien à des erreurs sur le bilan de masse plus faibles que le schéma NI standard pour une chimie cinétique, mais les erreurs sur le profil de concentration sont équivalentes pour les 2 schémas. De plus, le schéma Strang-splitting n'est absolument pas adapté à la description d'un système incluant une chimie à l'équilibre instantané. En effet, le principe d'un tels système est que l'équilibre chimique soit respecté à la fin de chaque pas de temps ; alors que le schéma Strang-splitting consiste en un premier demi-pas de temps de transport (soit une perturbation de l'équilibre chimique), suivi d'une résolution de l'opérateur de chimie (soit une mise à l'équilibre chimique), puis d'un second demi-pas de temps de transport (donc une perturbation de l'équilibre chimique).

2.3. RESOLUTION DE L'EQUILIBRE CHIMIQUE

2.3.1. METHODE CLASSIQUE DE NEWTON-RAPHSON

Les premiers travaux en modélisation des systèmes chimiques à l'équilibre [91, 179, 180, 200] ont été développés en utilisant la méthode de Newton-Rapshon pour minimiser un système algébrique non linéaire.

Nous avons montré précédemment (§ 1.4.2.2) qu'il est possible d'écrire tout système chimique à l'équilibre thermodynamique sous la forme d'un système constitué de Nc lois d'action de masse :

$$\{c_i\} = K_i \prod_{j=1}^{Nx} \{X_j\}^{a_{i,j}} \quad \forall i = 1,...Nc \tag{1.131}$$

et de Nx équations de conservation.

$$[T_j] = \sum_{i=1}^{Nc} b_{i,j} \cdot [c_i] \quad \forall j = 1,...Nx \tag{1.134}$$

La fermeture du système est assurée par la relation entre activité et concentration :

$$\{c_i\} = \gamma_i [c_i] \tag{1.97}$$

Il peut s'avérer nécessaire d'introduire des modifications à cette formulation pour inclure certains phénomènes chimiques comme la complexation de surface (§ 1.4.2.1.6)

2.3.1.1. Procédure de résolution

En substituant les lois d'action de masse dans les équations de conservation à l'aide des coefficients d'activité, on obtient un système de Nx équations à Nx inconnues :

$$[T_j] = \sum_{i=1}^{Nc} b_{i,j} \cdot \frac{K_i}{\gamma_i} \prod_{k=1}^{Nx} \left(\gamma_k [X_k]\right)^{a_{ik}} \quad \forall j = 1,...Nx \tag{2.1}$$

Notons que dans le système (2.1), les coefficients d'activité γ dépendent des concentrations des espèces chimiques, donc des concentrations en composants.

La résolution de ce système non linéaire par la méthode de Newton-Raphson implique une procédure itérative (on notera ^{n} l'itération en cours). Si pour un jeu de concentrations en composants $[X_k^n]$, l'équation de conservation (2.1) n'est pas respectée, on notera ici Y_j^n le résidu pour le composant j à l'itération n :

$$Y_j^n = -[T_j] + \sum_{i=1}^{Nc} b_{i,j} \cdot \frac{K_i}{\gamma_i} \prod_{k=1}^{Nx} \left(\gamma_k [X_k^n]\right)^{a_{ik}} \quad \forall j = 1,...Nx \tag{2.2}$$

La matrice jacobienne $\mathbf{Z}$ de ce système non linéaire est définie par :

$$Z_{j,k}^{n} = \frac{\partial Y_j^n}{\partial \left[X_k\right]} \tag{2.3}$$

Si l'on suppose que les coefficients d'activité sont constants, le calcul analytique donne :

$$\begin{aligned} Z_{j,k}^{n} &= \sum_{i=1}^{Nc} b_{i,j} \cdot \frac{K_i}{\gamma_i} \prod_{\substack{h=1 \\ h\neq k}}^{Nx} \left(\gamma_h \left[X_h^n\right]\right)^{a_{i,h}} \cdot a_{i,k} \cdot \gamma_k^{\,a_{i,h}} \left[X_k^n\right]^{a_{i,h}-1} \\ Z_{j,k}^{n} &= \sum_{i=1}^{Nc} b_{i,j} \cdot a_{i,k} \cdot \frac{1}{\gamma_i} \frac{K_i \prod_{h=1}^{Nx} \left(\gamma_h \left[X_h^n\right]\right)^{a_{i,h}}}{\left[X_k^n\right]} = \sum_{i=1}^{Nc} \frac{b_{i,j} \cdot a_{i,k}}{\gamma_i} \frac{\left[c_i^n\right]}{\left[X_k^n\right]} \end{aligned} \tag{2.4}$$

Il est également possible de définir la jacobienne par rapport à une variation des concentrations en composants $\Delta\left[X_k^n\right]$:

$$Z_{j,k}^{n} = \frac{\Delta Y_j^n}{\Delta\left[X_k^n\right]} = \frac{Y_j^{n+1} - Y_j^n}{\Delta\left[X_k^n\right]} \tag{2.5}$$

L'objectif du travail étant de trouver le vecteur de concentration $\left[X_k^{n+1}\right]$ tel que le résidu Y_j^{n+1} soit nul, on peut alors écrire :

$$Z_{j,k}^{n} \cdot \Delta\left[X_k^n\right] = -Y_j^n \tag{2.6}$$

L'inversion du système (2.6) donne le pas d'avancement de la méthode $\Delta\left[X_k^n\right]$, qui permet de calculer $\left[X_k^{n+1}\right]$ le nouveau vecteur des concentrations en composants.

$$\left[X_k^{n+1}\right] = \left[X_k^n\right] + \Delta\left[X_k^n\right] \tag{2.7}$$

Cette procédure, de (2.2) à (2.7), est répétée jusqu'à convergence, c'est-à-dire jusqu'à ce que le résidu Y_j^n soit inférieur à un critère prédéfini : ε_{NR} .

2.3.1.2. Convergence

Dans un formalisme mathématique, on trouve deux façons usuelles de présenter le critère de convergence de la méthode de Newton-Raphson.

(i) Une formulation en termes de pas d'avancement :

$$\left\| \left[\mathbf{X}^{n+1}\right] - \left[\mathbf{X}^{n}\right] \right\| \leq \varepsilon_{NR} \tag{2.8}$$

où $\| \ \|$ est souvent la norme vectorielle et $\left[\mathbf{X}^n\right]$ le vecteur des concentrations en composants.

(ii) Une formulation en terme de minimalisation du résidu :

$$\left\| \mathbf{Y}^n \right\| \leq \varepsilon_{NR} \tag{2.9}$$

Dans le cas de l'application de la méthode de Newton-Raphson au problème spécifique de calcul des équilibres chimiques, ces deux formulations ne sont pas adaptées. En effet, la formulation (2.8) ne permet pas de détecter un minimum local. La formulation (2.9) quant à elle conduit à exiger la même précision de résolution pour toutes les espèces chimiques, alors que nous sommes souvent confrontés à des différences de concentrations de plusieurs ordres de grandeur entre les éléments majeurs (Ca^{2+} , HCO_3^- , concentrations souvent supérieures à 10^{-2} mol/L) et les éléments traces (métaux lourds, pesticides, radioéléments, concentrations souvent inférieures à 10^{-6} mol/L).

2.3.2. RESTRICTION DU DOMAINE CHIMIQUE ET FRACTIONS CONTINUES POSITIVES

Il est reconnu que la méthode de Newton Raphson peut ne pas converger, si le jeu initial de composants proposé est loin de la solution. Dans l'optique d'une modélisation de transport réactif, cet échec de du processus de minimisation peut faire échouer tout le calcul. Nous avons donc travaillé sur cette problématique afin de proposer des solutions.

Dans une première analyse, nous avons détecté que les difficultés proviennent d'une matrice jacobienne singulière ou quasi-singulière, traduisant une pente nulle ou quasi nulle de la fonction résidu.

Lorsque la matrice jacobienne est singulière, il n'est pas possible d'inverser le système (2.6) et la méthode échoue. La solution apportée est de faire appel à une méthode d'ordre zéro, ne nécessitant pas le calcul de la pente de la fonction résidu. Après avoir testé et éliminé les méthodes d'ordre zéro les plus classiques (Picard et point fixe), deux méthodes ont été retenues : la méthode du simplex et la méthode des fractions continues. Sous sa forme originale, la méthode des fractions continue s'est révélée trop rigide et ne permet pas la prise en compte de coefficients stœchiométriques négatifs. Nous avons proposé une modification de cette méthode, appelée méthode des fractions continues positives, qui possède désormais la flexibilité souhaitée.

Lorsque la matrice est quasi-singulière, le pas d'avancement $\Delta\left[\mathbf{X}^n\right]$ obtenu est alors très grand. Les concentrations en composant à l'itération suivante $\left[\mathbf{X}^{n+1}\right]$ n'ont alors plus aucune signification physique. Nous avons proposé de limiter les valeurs possibles pour les composants en définissant des bornes construites sur la base de critères chimiques : le domaine chimiquement autorisé.

Nous avons montré qu'en combinant ces deux nouveaux éléments à la méthode classique de Newton-Raphson, on obtient un algorithme à la fois plus robuste et plus rapide que les méthodes existantes. Cet algorithme a été implanté dans le code de spéciation SPECY.

L'ensemble de ces méthodes ont été testées et comparées. Les critères de comparaison retenus ont été la robustesse de la méthode et la vitesse de convergence. Deux systèmes chimiques ont été choisis pour les tests en raison de la différence de comportement des méthodes qu'ils induisent. Le test de l'acide gallique est proposé par Brassard et Bodurta [201]. Ce test présente un piège en boucle qui bloque les méthodes basées sur la pente : les directions de descente successives renvoient le processus de recherche au même point, comme s'il rebondissait contre les parois d'une vallée. Le test de la pyrite est proposé par J. Van der Lee (communication personnelle). Ce test présente de très grands écarts entre les concentrations des espèces et des composants. La concentration initiale choisie pour le composant $\left[O_2\right]$ est de 10^{-3} mol/L (situation proche de conditions aérobies) alors que la concentration à l'équilibre est de 1,10 10^{-73} mol/L.

Notons que cette concentration ne représente alors pas un nombre de molécule d'oxygène (une molécule pour 10^{37} km^3) mais un potentiel redox très bas, car nous avons choisi d'utiliser l'oxygène comme accepteur/donneur d'électron. Les autres composants, H^+, Fe^{2+} et SO_4^{2-} ont des concentrations à l'équilibre entre 10^{-9} et 10^{-7} mol/L. Les concentrations des espèces dissoutes sont comprises entre 10^{-6} et 10^{-20} mol/L car nous avons décidé de négliger les concentrations des espèces inférieures à 10^{-23} mol/L (cela correspond à une molécule par litre).

Les tests montrent bien l'intérêt de l'algorithme combiné, qui permet de réduire le temps de calcul et d'éliminer totalement les problèmes de non convergence. Ce travail a été publié dans *AIChE Journal* [92] et peut être trouvé en Annexe 3).

2.3.3. FORMULATION EN LOGARITHME ET CONDITIONNEMENT DE LA JACOBIENNE

Lors de la conception de l'exercice de comparaison de codes pour le GdR MoMaS, l'objectif était de proposer un cas test permettant de mettre à l'épreuve les différents outils numériques existants. Le niveau de non linéarité et de couplage du système chimique a ainsi été augmenté jusqu'à mettre en échec l'algorithme de SPECY. Par la suite, il a fallu développer de nouveaux outils pour pouvoir fournir des solutions à l'exercice proposé.

2.3.3.1. Changement de variables : composants en logarithme d'activité

2.3.3.1.1. Présentation du concept

La solution adoptée consiste en un changement de variables effectué dans le système (2.2). Au lieu de travailler sur les concentrations en composants $\left[X_j\right]$, nous avons écrit ce système en logarithme de l'activité des composants ξ_j [99].

$$\xi_j = ln\left\{X_j\right\} = ln\left(\gamma_j\left[X_j\right]\right) \tag{2.10}$$

Ou, sous un formalisme matriciel :

$$\xi = \ln\ diag\ \gamma_X\ \mathbf{X}\ = diag\ \ln\gamma_X\ + \ln\ \mathbf{X} \tag{2.11}$$

On peut alors reformuler les relations chimiques de la façon suivante, pour les lois d'action de masse :

$$\ln\mathbf{C} = \ln\mathbf{K} - diag\ \ln\gamma_C\ + \mathbf{A}\cdot\xi \tag{2.12}$$

Et pour les équations de conservation :

$$\mathbf{T} = {}^T\mathbf{B}\cdot\mathbf{C} \tag{2.13}$$

Le système (2.2) s'écrit alors en fonction de la variable ξ :

$$Y_j^n = -\left[T_j\right] + \sum_{i=1}^{Nc} b_{i,j} \cdot \frac{K_i}{\gamma_i} \prod_{k=1}^{Nx} \left(exp\left(\xi_k^n\right)\right)^{a_{i,k}} \quad \forall j = 1,...Nx \tag{2.14}$$

Ce qui peut se reformuler de la façon suivante :

$$Y_j^n = -\left[T_j\right] + \sum_{i=1}^{Nc} b_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot exp\left(\sum_{k=1}^{Nx} a_{i,k} \cdot \xi_k^n\right) \quad \forall j = 1,...Nx \tag{2.15}$$

Ou sous forme matricielle :

$$\mathbf{Y} = -\mathbf{T} + {}^{T}\mathbf{B} \cdot \exp\ \ln \mathbf{K} - diag\ \ln \gamma_C\ + \mathbf{A} \cdot \xi \tag{2.16}$$

Sous ce formalisme, le calcul analytique de la matrice jacobienne $\tilde{\mathbf{Z}}$ donne, si l'on néglige les variations du coefficient d'activité γ_i :

$$\tilde{Z}_{j,k}^n = \frac{\partial Y_j^n}{\partial \xi_k} = \sum_{i=1}^{Nc} b_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot a_{i,k} \cdot exp\left(\sum_{h=1}^{Nx} a_{i,h} \cdot \xi_h^n\right) = \sum_{i=1}^{Nc} b_{i,j} \cdot a_{i,k} \cdot \left[c_i^n\right] \quad \forall j,k = \left(1,...Nx\right)^2 \tag{2.17}$$

Soit :

$$\tilde{\mathbf{Z}} = {}^{T}\mathbf{B} \cdot diag\ \mathbf{C}\ \cdot \mathbf{A} \tag{2.18}$$

Le système linéaire (2.6) à inverser s'écrit alors sous une forme similaire (2.19) :

$$\tilde{\mathbf{Z}}^n \cdot \Delta\left[\xi^n\right] = -\mathbf{Y}^n \tag{2.19}$$

En comparant la matrice jacobienne $\tilde{\mathbf{Z}}$ obtenue en (2.17) avec $\mathbf{Z}$ issue de la formulation initiale (2.4) on constate deux points :

i) Le terme contenant initialement $\frac{[c_i]}{[X_k]}$ a été remplacé par $[c_i]$. Ainsi, les termes de la matrice $\tilde{\mathbf{Z}}$ sont tous du même ordre de grandeur que les concentrations (de 10^{-15} à 10^{-1} mol/L en pratique). En revanche, certains termes de la matrice $\mathbf{Z}$ peuvent être du même ordre de grandeur que $\frac{[c_i]}{[X_k]}$, c'est-à-dire selon le cas $\frac{10^{-15}}{10^{-1}}$ ou $\frac{10^{-1}}{10^{-15}}$.

ii) La matrice $\mathbf{Z}$ est remplie par blocs, selon les coefficients stœchiométriques des composants, de même que la matrice $\tilde{\mathbf{Z}}$. Mais la matrice $\tilde{\mathbf{Z}}$ est en plus symétrique dans la plupart des cas, car les coefficients stœchiométriques $a_{i,j}$ et $b_{i,j}$ sont très souvent identiques.

Par ce moyen, on obtient alors une matrice jacobienne mieux conditionnée et symétrique dans la plupart des cas.

De plus, lorsque l'on travaille avec les concentrations en composants $\left[X_j\right]$, il arrive parfois d'un pas d'avancement $\Delta\left[X_k^n\right]$ négatif conduise à des concentrations négatives, ce qui oblige à modifier arbitrairement le nouveau vecteur des composants. En travaillant en logarithme de l'activité ξ_j, un pas d'avancement négatif ne conduit jamais à une solution non physique. Il n'est donc plus nécessaire d'imposer la borne inférieure du domaine chimiquement autorisé.

Ces améliorations ont été testées dans le cadre du benchmark Reactive Transport du GdR MoMaS [84, 97-99]. La présentation d'une comparaison entre le code dans sa première version et celui obtenu suite à ces améliorations n'était pas pertinente dans le cadre de ce benchmark. Soulignons néanmoins que nous n'avons pas pu résoudre le problème proposé avec le code dans sa version initiale dans des temps de calcul raisonnables (j'ai abandonné lorsque le code a bloqué après 3 semaines de calcul sans avoir passé l'étape d'injection). Avec ces améliorations, il est possible de résoudre le test du benchmark, avec un pas d'espace et de temps assez grand pour avoir des temps de calcul court (environ une demi-heure). Cependant, il est nécessaire d'utiliser un maillage fin et un pas de temps petit pour obtenir une bonne résolution.

Il est à noter que des travaux récents [202] conduisent à des résultats opposés, car ces auteurs recommandent l'utilisation d'une formulation en concentration plutôt qu'en logarithme d'activité. Ils montrent, entre autre, que pour une formulation en logarithme, la résolution du système conduit à une matrice jacobienne singulière si la concentration totale en composant tend vers 0. Ces résultats ne sont pas incompatibles avec ceux présentés en Annexe 4 car les tests effectués [203] n'incluent pas de tests avec des concentrations totales nulles ou quasi nulles, et les travaux de Erhel et Sabit [202] portent sur un modèle de transport réactif par approche globale. Ainsi la matrice jacobienne calculée par ces auteurs inclus des termes de transport. Cependant, ce point mérite d'être étudié plus avant et il faudra entre autre reprendre une partie des tests utilisés dans le travail [203] en les modifiant pour voir comment évoluent les matrices jacobiennes lorsque les concentrations totales tendent vers 0. Suivant le cas, on pourrait essayer de déterminer un critère pour choisir à priori la formulation la plus pertinente et basculer d'une formulation en concentration vers une formulation en logarithme d'activité selon les conditions.

2.3.3.1.2. Prise en compte des cas particuliers

Cependant, ce nouveau formalisme ne permet plus de passer rapidement d'une résolution par la méthode des fractions continues positive à la méthode de Newton-Raphson. Il est à priori nécessaire de recalculer des concentrations en composant, ce qui peut s'avérer problématique d'un point de vue informatique, surtout lorsque ξ_j est négatif avec une grande valeur absolue ($\xi_j \leq -400$). De plus, si l'intégration de phénomènes chimiques particuliers comme ceux de précipitation-dissolution ou de complexation de surface se font aisément dans le formalisme en concentration [92], cela n'est pas forcément le cas dans un formalisme en logarithme d'activité.

Précipitation-dissolution

L'algorithme général de prise en compte de la précipitation-dissolution n'a pas été modifié par rapport à ce qui a été fait précédemment [92]. La loi d'action de masse **(1.139)** se réécrit sans difficulté en logarithme d'activité :

$$1 = K_i \prod_j exp\left(\xi_j\right)^{a_{i,j}} \text{ soit } 0 = ln\left(K_i\right) + \sum_j a_{i,j}\,\xi_j \tag{2.20}$$

On définit alors le résidu Yp_i^n à l'itération n pour le précipité i de la façon suivante :

$$Yp_{Nx+i}^n = ln\left(K_i\right) + \sum_j a_{i,j}\,\xi_j^{\,n} \quad \forall i = 1,...NcP \tag{2.21}$$

et le résidu (2.15) est modifié pour inclure la quantité de précipité formé :

$$Y_j^n = -\left[T_j\right] + \sum_{i=1}^{Nc-NcP} b_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot exp\left(\sum_{k=1}^{Nx} a_{i,k} \cdot \xi_k^n\right) + \sum_{i=1}^{NcP} b_{i,j} \cdot \left[cp_i^n\right] \quad \forall j = 1,...Nx \tag{2.22}$$

Le terme générique de la matrice jacobienne s'écrit dans ce cas :

$$\tilde{Z}_{Nx+i,k}^n = \frac{\partial Yp_i^n}{\partial \xi_k} = a_{i,k} \quad \forall i = 1,...NcP, \forall k = 1,...Nx \text{ et } \tilde{Z}_{i,Nx+k}^n = \frac{\partial Y_i^n}{\partial\left[cp_k^n\right]} = b_{i,k} \quad \forall k = 1,...NcP, \forall i = 1,...Nx \tag{2.23}$$

Complexation de surface

Le composant électrostatique X_Ψ défini par l'équation **(1.143)** peut se redéfinir en logarithme d'activité ξ_Ψ :

$$\xi_\Psi = -\frac{F}{RT^\circ} \cdot \Psi \tag{2.24}$$

Le résidu sur l'équation de conservation des sites de surface Ts_j s'écrit :

$$Y_j^n = \sum_{i=1}^{NcS} b_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot exp\left(\xi_\Psi^n + \sum_{k=1}^{Nx} a_{i,k}\xi_k^n\right) = \sum_{i=1}^{NcS} b_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot exp\left(\sum_{\substack{k=1 \\ \Psi inclu}}^{Nx} a_{i,k}\xi_k^n\right) \tag{2.25}$$

On constate donc que sur l'équation (2.25), le composant potentiel ξ_Ψ se comporte comme un composant classique, avec un coefficient stœchiométrique $a_{i,\Psi}$ égal à un.

Le résidu sur l'équation de conservation des charges électrostatiques s'écrit lui selon l'équation (2.26) avec la particularité que le total T_Ψ dépend du composant potentiel X_Ψ selon l'équation **(1.147)** :

$$Y_\Psi^n = -\left[T_\Psi^n\right] + \sum_i z_i\left[c_i^n\right] = -\frac{S \cdot \rho_S}{F} \cdot f_{CS}\left(-\frac{RT^\circ}{F} ln\left\{X_\Psi^n\right\}\right) + \sum_i b_{i,\Psi}\left[c_i^n\right] \tag{2.26}$$

En incluant les composants ξ :

$$Y_{\Psi}^{n} = -\frac{S \cdot \rho_S}{F} \cdot f_{CS}\left(-\frac{RT^{\circ}}{F}\xi_{\Psi}^{n}\right) + \sum_i b_{i,\Psi} \cdot \frac{K_i}{\gamma_i} \cdot exp\left(\sum_{k=1}^{Nx} a_{i,k}\xi_k^n\right) \tag{2.27}$$

Pour le composant potentiel de surface, il faut donc modifier le terme de la matrice jacobienne relatif à ce composant.

$$\begin{aligned} \tilde{Z}_{\Psi,j}^{n} &= \frac{\partial Y_{\Psi}^{n}}{\partial \xi_j} = \sum_i b_{i,\Psi} \cdot \frac{K_i}{\gamma_i} \cdot \frac{\partial}{\partial \xi_j}\left[exp\left(\sum_{k=1}^{Nx} a_{i,k}\xi_k^n\right)\right] = \sum_i b_{i,\Psi} \cdot a_{i,j} \cdot \left[c_i^n\right] \\ \tilde{Z}_{j,\Psi}^{n} &= \frac{\partial Y_{j}^{n}}{\partial \xi_{\Psi}} = \sum_i b_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot \frac{\partial}{\partial \xi_{\Psi}}\left[exp\left(\sum_{k=1}^{Nx} a_{i,k}\xi_k^n\right)\right] = \sum_i b_{i,j} \cdot a_{i,\Psi} \cdot \left[c_i^n\right] \end{aligned} \tag{2.28}$$

On constate que pour les termes hors diagonaux, la matrice se calcule de façon identique aux termes génériques (2.17). Les termes diagonaux incluent une part provenant de la dérivée de la quantité totale de charges :

$$\begin{aligned} \tilde{Z}_{\Psi,\Psi}^{n} &= \frac{\partial Y_{\Psi}^{n}}{\partial \xi_{\Psi}} = -\frac{S \cdot \rho_S}{F} \cdot \frac{\partial}{\partial \xi_{\Psi}}\left[f_{CS}\left(-\frac{RT^{\circ}}{F}\xi_{\Psi}^{n}\right)\right] + \sum_i z_i \cdot \frac{K_i}{\gamma_i} \cdot \frac{\partial}{\partial \xi_{\Psi}}\left[exp\left(\sum_{k=1}^{Nx} a_{i,k}\xi_k^n\right)\right] \\ \tilde{Z}_{\Psi,\Psi}^{n} &= -\frac{S \cdot \rho_S}{F} \cdot \frac{\partial}{\partial \xi_{\Psi}}\left[f_{CS}\left(-\frac{RT^{\circ}}{F}\xi_{\Psi}^{n}\right)\right] + \sum_i z_i \cdot a_{i,j} \cdot \left[c_i^n\right] \end{aligned} \tag{2.29}$$

On constate ainsi que la matrice jacobienne obtenue en choisissant les ξ_j comme composants conduit à une matrice symétrique si les coefficients stœchiométriques $a_{i,j}$ et $b_{i,j}$ sont égaux, et ceci même en incluant les phénomènes de précipitation-dissolution ou de complexation de surface.

Solutions solides

Il est possible de décrire la formation d'une solution solide en exprimant les concentrations de chaque pôle en concentration ou en logarithme de concentration. Cependant, la disparition d'une solution solide impose une expression en concentration. En effet, l'algorithme de résolution classique commence par une hypothèse : on suppose que la solution solide se forme (ou non). Si la résolution des équations de conservation et de loi d'action de masse conduit à des concentrations négatives pour l'un des pôles, alors l'hypothèse faite se révèle erronée et la résolution doit être recommencée avec l'hypothèse contraire.

La loi d'action de masse **(1.140)**, décrivant la formation d'un pôle Pi au sein de la solution solide S comprenant N_{Ps} pôles, s'écrirait alors :

$$0 \leq \ln\ K_i\ + \sum_j \xi_j^{a_{i,j}} - \ln\left(\frac{n_{Pi}}{\sum_{Pk=1}^{N_{Ps}} n_{Pk}}\right) \tag{2.30}$$

Cependant, au cours du processus de minimisation, il se peut que le terme $\sum_{Pk=1}^{N_{Ps}} n_{Pk}$ devienne négatif, ce qui conduirait à un arrêt du code. Afin d'éviter cela, il est nécessaire de conserver un formalisme plus proche de l'écriture classique qui donne pour la loi d'action de masse :

$$1 \leq K_i \cdot \exp\left(\sum_j \xi_j^{a_{i,j}}\right) \cdot \frac{\sum_{Pk=1}^{N_{Ps}} n_{Pk}}{n_{Pi}} \tag{2.31}$$

Il est ainsi possible qu'au cours de la minimisation, le terme $\sum_{Pk=1}^{N_{Ps}} n_{Pk}$ devienne négatif, mais il est également garanti que ce terme soit positif une fois la convergence obtenue.

Nous avons ainsi un ensemble de $Npss$ (Nombre de pôles des solutions solides) variables primaires qui s'ajoutent aux Nx composants existants. Le résidu sur les équations de conservation de matière pour les composants s'écrit alors :

$$Y_j^n = -\left[T_j\right] + \sum_{i=1}^{Nc-NcP} b_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot \exp\left(\sum_{k=1}^{Nx} a_{i,k} \cdot \xi_k^n\right) + \sum_{i=1}^{NcP} b_{i,j} \cdot \left[cp_i^n\right] + \sum_{Pi=1}^{Npss} bss_{Pi,j} \cdot n_{Pi}^n \quad \forall j = 1,...Nx \tag{2.32}$$

Le résidu sur la loi d'action de masse relative à la formation du pôle Pi d'une solution solide donne :

$$Y_{SSi}^n = -1 + K_i \cdot \exp\left(\sum_j \xi_j^{n^{a_{i,j}}}\right) \cdot \frac{\sum_{k=1}^{N_{Ps}} n_{Pk}}{n^n{}_{Pi}} \tag{2.33}$$

Le terme de la matrice jacobienne relatif à la formation du pôle Pi dans le résidu en bilan de masse du composant j s'écrit :

$$Z_{j,Pi}^n = \frac{\partial Y_j^n}{\partial n_{Pi}^n} = bss_{Pi,j} \tag{2.34}$$

Le terme de la matrice jacobienne relatif à la formation du pôle Pi dans le résidu de loi d'action de masse de la formation du pôle Pi donne :

$$Z_{Pi,Pi}^n = \frac{\partial Y_{SSi}^n}{\partial n^n{}_{Pi}} = K_i \cdot \exp\left(\sum_j \xi_j^{n^{a_{i,j}}}\right) \frac{1}{n^n{}_{Pi}} \cdot \left(1 - \frac{\sum_{k=1}^{N_{Ps}} n_{Pk}}{n^n{}_{Pi}}\right) \tag{2.35}$$

Le terme de la matrice jacobienne relatif à la formation du pôle Pm dans le résidu de loi d'action de masse de la formation du pôle Pi s'écrit :

$$Z_{Pi,Pm}^n = \frac{\partial Y_{SSi}^n}{\partial n^n{}_{Pm}} = K_i \cdot \exp\left(\sum_j \xi_j^{n^{a_{i,j}}}\right) \frac{1}{n^n{}_{Pi}} \tag{2.36}$$

Fractions continues positives

Telle qu'elle est présentée dans les travaux précédent [92], la méthode des fractions continues positives exprime le résidu sur les équations de conservation de la matière sous la forme d'une différence entre la somme des réactifs et la somme des produits :

$$Sum_{react,j} = \sum_{\substack{i \\ b_{i,j}<0}} b_{i,j}\left[c_i\right] \text{ et } Sum_{product,j} = T_j + \sum_{\substack{i \\ b_{i,j}>0}} b_{i,j}\left[c_i\right] \tag{2.37}$$

Pour un jeu de concentrations en composant, on définit la valeur suivante en fonction du rapport entre la somme des réactifs et celle des produits.

$$X_j^{n+1} = X_j^n \cdot \left(\frac{Sum_{react,j}}{Sum_{product,j}} \right)^{1/a_{i^\circ,j}} \tag{2.38}$$

Cependant, la méthode des Fractions Continues Positives (comme toutes les méthodes d'ordre zéro) ne prennent pas en compte l'impact sur la fonction Y_j d'une modification de la valeur X_k (à la différence des méthodes d'ordre un). La modification simultanée des valeurs de tous les composants entraîne des oscillations. L'introduction d'une moyenne pondérée permet de restreindre les oscillations et d'assurer la convergence dans un maximum de cas.

$$X_j^{n+1} = \theta \cdot X_j^n \cdot \left(\frac{Sum_{react,j}}{Sum_{product,j}} \right)^{1/a_{i^\circ,j}} + (1-\theta) \cdot X_j^n \tag{2.39}$$

avec θ compris entre 0 et 1.

Une valeur de θ forte en début de recherche (θ = 0,9) permet de se rapprocher rapidement de la solution. En effet, ceci revient à diviser (resp. multiplier) X^n par 10 si sa valeur est très grande (resp. très petite) devant la valeur d'équilibre. En fin de recherche, il est important de limiter les oscillations et une valeur de θ plus faible (θ = 0,1) permet d'assurer la stabilité de la méthode. Nous avons donc montré qu'il était intéressant de construire θ comme un paramètre adaptatif qui décroît au fur et à mesure que l'on s'approche de la solution.

$$\begin{aligned} &\text{if } \left(Sum_j^{reac} < Sum_j^{prod}\right) \text{ then } \left(\theta_j = 0.9 - \frac{Sum_j^{reac}}{Sum_j^{prod}} \cdot 0.8\right) \\ &\text{if } \left(Sum_j^{reac} > Sum_j^{prod}\right) \text{ then } \left(\theta_j = 0.9 - \frac{Sum_j^{prod}}{Sum_j^{reac}} \cdot 0.8\right) \end{aligned} \tag{2.40}$$

L'utilisation d'une formulation en logarithme de l'activité des composants nécessite une réécriture de cette méthode des fractions continues positives. La relation (2.38) se prête bien à une réécriture en logarithme, mais pas la relation opérationnelle (2.39). Cependant, la relation (2.39) n'est qu'une façon d'introduire un terme de relaxation dans la méthode. Il est tout à fait possible d'introduire cette relaxation de façon différente. Ecrite en fonction du logarithme de l'activité des composants, la relation (2.38) donne :

$$\xi_j^{n+1} = \xi_j^n + \frac{1}{a_{i^\circ,j}}\left[ln\left(Sum_{react,j} \right) - ln\left(Sum_{product,j} \right) \right] \tag{2.41}$$

Il est alors très simple d'introduire un terme de relaxation dans la relation (2.41) :

$$\xi_j^{n+1} = (1-\theta)\xi_j^n + \frac{\theta}{a_{i^\circ,j}}\left[ln\left(Sum_{react,j} \right) - ln\left(Sum_{product,j} \right) \right] \tag{2.42}$$

La relation (2.40) doit alors être modifiée pour conserver les performances de la méthode des fractions continues positives :

$$\begin{aligned} &\text{if } \left(Sum_j^{reac} < Sum_j^{prod} \right) \text{ then } \left(\theta_j = 0.1 - \frac{Sum_j^{reac}}{Sum_j^{prod}} \cdot 0.08 \right) \\ &\text{if } \left(Sum_j^{reac} > Sum_j^{prod} \right) \text{ then } \left(\theta_j = 0.1 - \frac{Sum_j^{prod}}{Sum_j^{reac}} \cdot 0.08 \right) \end{aligned} \tag{2.43}$$

2.3.3.2. Structure et conditionnement de la jacobienne

La matrice jacobienne associée au calcul d'un équilibre chimique possède une structure propre, similaire pour tous les systèmes. Issue de la structure du tableau des équilibres Tableau 1-10 et des interactions possibles ou non entre les différentes phases, la structure de la matrice jacobienne est plutôt creuse, carrée par blocs, comme cela peut se voir sur le Tableau 2-1.

Dans l'optique d'un modèle numérique de transport réactif par séparation d'opérateurs, il est nécessaire de résoudre le problème de chimie à l'équilibre au moins une fois (plusieurs fois dans le cas de schémas itératifs) par maille (ou nœud selon la discrétisation spatiale) et par pas de temps. Cela implique donc la résolution d'un très grand nombre de systèmes linéaires (2.19).

Ces système linéaires ont comme propriété d'être de relativement petite taille : 4 x 4 pour des applications de laboratoire simples, jusqu'à 20 x 20 pour des cas réels extrêmement complexes. On peut envisager dans un avenir proche des cas allant jusqu'à 50 composants, mais la dimension du problème reste relativement faible. D'autre part, nous avons déjà souligné que le conditionnement de la matrice jacobienne est relativement mauvais, et ce même après une reformulation en logarithme d'activité des composants.

Pour la première fois, nous avons étudié de façon systématique le conditionnement des matrices jacobiennes issues de la modélisation des systèmes chimiques à l'équilibre (voir Annexe 4) ainsi que de nombreuses méthodes directes (6) et itératives (6) de résolution des systèmes linéaires. Une telle étude n'avait jamais été réalisée pour des matrices de petites tailles, mal conditionnées. Nous proposons une compilation de 10 cas-test de chimie à l'équilibre, de dimensions variables (de 3 à 52 composants, de 7 à 780 espèces chimique), conduisant à des matrices jacobiennes dont les conditionnements varient de $10^{0,61}$ à $10^{213,9}$. Dans ce travail [203], nous avons mis en évidence plusieurs points importants :

- Le conditionnement de ces petites matrices peut être extrêmement élevé, rendant la résolution du système linéaire difficile. La non-convergence de la méthode de Newton-Raphson dans certains cas peut s'expliquer par l'impossibilité de déterminer une direction de descente adaptée.
- Il existe, pour les grands nombres de conditionnement, une relation entre la norme de l'erreur $\|Y\|$ et le conditionnement de la matrice jacobienne. Nous n'avons pas réussi à justifier

mathématiquement l'existence de cette relation. Cependant, ceci permet, d'un point de vue opérationnel, d'estimer le conditionnement de la matrice jacobienne pour un coût en temps de calcul minime.

- Certaines méthodes de résolution (Décomposition LU et GMRES) sont plus rapides et moins sensibles au conditionnement que les autres méthodes testées. Lorsque le conditionnement de la matrice est fort, nous recommandons donc l'usage de ces méthodes.

Tableau 2-1 : structure générique de la matrice jacobienne des systèmes chimiques

<table>
<tr><td></td><td>$\frac{\partial}{\partial \xi g}$</td><td>$\frac{\partial}{\partial \xi a}$</td><td>$\frac{\partial}{\partial \xi e}$</td><td>$\frac{\partial}{\partial \xi s}$</td><td>$\frac{\partial}{\partial \xi \psi}$</td><td>$\frac{\partial}{\partial [cp]}$</td></tr>
<tr><td>$\frac{\partial Yg}{\partial}$</td><td colspan="3" rowspan="4">$\sum_{i=1}^{Nc} b_{i,j} \cdot a_{i,k} \cdot \left[c_i^n\right]$</td><td colspan="2">$\sum_i b_{i,j} \cdot a_{i,\Psi} \cdot \left[c_i^n\right]$</td><td>$b_{i,k}$</td></tr>
<tr><td>$\frac{\partial Ya}{\partial}$</td><td colspan="3"></td><td colspan="2"></td><td></td></tr>
<tr><td>$\frac{\partial Ye}{\partial}$</td><td colspan="2"></td><td colspan="3">0</td><td rowspan="3">0</td></tr>
<tr><td>$\frac{\partial Ys}{\partial}$</td><td colspan="2" rowspan="2"></td><td>$\sum_{i=1}^{Nc} b_{i,j} \cdot a_{i,k} \cdot \left[c_i^n\right]$</td><td>$\sum_i b_{i,j} \cdot a_{i,\Psi} \cdot \left[c_i^n\right]$</td><td></td></tr>
<tr><td>$\frac{\partial Y\psi}{\partial}$</td><td>$\sum_i b_{i,\Psi} \cdot a_{i,j} \cdot \left[c_i^n\right]$</td><td>0</td><td>$\sum_i b_{i,\Psi} \cdot a_{i,j} \cdot \left[c_i^n\right]$</td><td>$-\frac{S\rho_s}{F} \cdot \frac{\partial}{\partial \xi_\Psi}\left[f_{cs}\left(-\frac{RT^\circ}{F}\xi_\Psi^n\right)\right] + \sum_i z_i \cdot a_{i,j} \cdot \left[c_i^n\right]$</td><td></td></tr>
<tr><td>$\frac{\partial Yp}{\partial}$</td><td>$a_{i,k}$</td><td colspan="4">0</td><td>0</td></tr>
</table>

2.3.3.3. Pré-conditionnement de la matrice jacobienne

Il est classiquement reconnu [204-206] que l'on ne peut obtenir une solution avec n décimales exactes à un système linéaire de conditionnement p que si l'on dispose d'une capacité de calcul à p-n chiffres significatifs. Or les conditionnements mis en évidence dans le travail de H. Machat [203] ne permettent pas d'envisager une solution correcte avec les moyens de calculs actuels. De nombreux outils mathématiques existent pour réduire le conditionnement d'un système linéaire. Dans le cadre de la thèse de M. Marinonni, nous avons étudié la possibilité d'utiliser de telles méthodes [207]. Le détail de ce travail est présenté en Annexe 5.

Le principe des techniques de préconditionnement est de multiplier un système linéaire une ou deux matrices simples (diagonales) afin de réduire le conditionnement du système obtenu. Ainsi, le système (2.19) sera remplacé par le système (2.44) :

$$\mathbf{D}_1 \cdot \tilde{\mathbf{Z}}^n \cdot \mathbf{D}_2 \cdot \Delta\left[\boldsymbol{\xi}^n\right] = -\mathbf{D}_1 \cdot \mathbf{Y}^n \qquad (2.44)$$

où $\mathbf{D}_1$ et $\mathbf{D}_2$ sont les deux matrices de préconditionnement et $\Delta\left[\xi^n\right] = \mathbf{D}_2^{-1} \cdot \Delta\left[\xi^n\right]$. Les systèmes (2.19) et (2.44) sont analytiquement équivalents, mais un choix judicieux des matrices $\mathbf{D}_1$ et $\mathbf{D}_2$ conduira à un meilleur conditionnement du système (2.44). Il est cependant reconnu [204] que l'efficacité de ces méthodes est fortement problème-dépendant et que seul un choix adapté des matrices $\mathbf{D}_1$ et $\mathbf{D}_2$ conduit à une diminution du conditionnement du système linéaire.

Nous avons testé plusieurs combinaisons de matrices $\mathbf{D}_1$ et $\mathbf{D}_2$ pour examiner les différentes performances de ces préconditionnements.

(i) Préconditionnement **RId** (Row Identity). La matrice $\mathbf{D}_2$ est la matrice identité, ce qui restreint le préconditionnement aux lignes de la matrice jacobienne. Si l'on note $\mathbf{a}_i$ le vecteur formé de la ligne i de la matrice $\tilde{\mathbf{Z}}^n$, on définit ici la matrice $\mathbf{D}_1$ par la relation (2.45).

$$D_{1i,i} = \left(\left\|\mathbf{a}_i\right\|_\infty\right)^{-1} \tag{2.45}$$

(ii) Préconditionnement **DId** (Diagonal Identity). La matrice $\mathbf{D}_2$ est la matrice identité et la matrice $\mathbf{D}_1$ est formée des termes diagonaux de la matrice $\tilde{\mathbf{Z}}^n$:

$$D_{1i,i} = \left(\tilde{\mathbf{Z}}^n_{i,i}\right)^{-1} \tag{2.46}$$

(iii) Préconditionnement **sDsD** (square Diagonal). Les matrices $\mathbf{D}_1$ et $\mathbf{D}_2$ sont définies par les racines carrées des termes diagonaux de la matrice $\tilde{\mathbf{Z}}^n$ (2.47). Ce préconditionnement est possible dans la plupart des cas chimiques car les matrices stœchiométriques relatives aux lois d'action de masse ($\mathbf{A}$) et aux équations de conservation ($\mathbf{B}$) sont très souvent identiques. De ce fait, la relation (2.17) garantit la positivité des termes diagonaux de $\tilde{\mathbf{Z}}^n$ pour peu que les concentrations des espèces $\left[c_i^n\right]$ soient positives, ce qui est garanti par la relation (2.12).

$$D_{1,i,i} = D_{2,i,i} = \sqrt{\tilde{Z}^n_{i,i}}^{-1} \tag{2.47}$$

(iv) Préconditionnement **MEq** (Matrix Equilibration). Il s'agit ici d'une procédure itérative, devant conduire à un conditionnement minimal de la matrice $\tilde{\mathbf{Z}}^n$. En notant $\mathbf{a}_i$ le vecteur formé de la ligne i et $\mathbf{c}_i$ le vecteur formé de la colonne i de la matrice $\tilde{\mathbf{Z}}^n$; les matrices $\mathbf{D}_1$ et $\mathbf{D}_2$ sont définies par les normes des vecteurs $\mathbf{a}_i$ et $\mathbf{c}_i$ (2.48). Knight et al. [208] recommandent 5 itérations de cette procédure pour obtenir un bon préconditionnement.

$$\begin{aligned} D_{1,i,i} &= \left(\sqrt{\left\|\mathbf{a}_i\right\|_\infty}\right)^{-1} \\ D_{2,i,i} &= \left(\sqrt{\left\|\mathbf{c}_i\right\|_\infty}\right)^{-1} \end{aligned} \tag{2.48}$$

Ces procédures de préconditionnement ont été testées sur plusieurs cas tests chimiques, pour un grand nombre de vecteur de composants initiaux. Le nombre d'itérations de Newton nécessaire pour obtenir la convergence a été comparé pour un algorithme sans préconditionnement et pour chacun de ces préconditionnements. Nous avons également testé l'utilisation systématique de la méthode FCP comme initialisation de la méthode de Newton-Raphson.

On montre que les techniques **sDsD** et **MEq** sont globalement plus efficaces que les autres, mais ces résultats restent très cas-dépendants. On distingue 3 cas de figures :

(i) Le conditionnement initial de la matrice $\tilde{\mathbf{Z}}^n$ est bas, et l'inversion du système (2.19) est suffisamment précise pour assurer la convergence rapide de l'algorithme de Newton. Dans ce cas, les techniques de préconditionnement sont inutiles et n'apportent aucun gain.

(ii) La matrice initiale $\tilde{\mathbf{Z}}^n$ a un conditionnement intermédiaire. L'inversion du système (2.19) n'est pas suffisamment précise ; et les techniques de préconditionnement permettent de réduire le conditionnement pour assurer une résolution précise du système (2.44). Dans ces cas intermédiaires, le préconditionnement apporte un gain appréciable en termes de convergence.

(iii) Lorsque le conditionnement initial de la matrice $\tilde{\mathbf{Z}}^n$ est trop grand, même les techniques de préconditionnement ne permettent pas d'obtenir une résolution précise du système (2.44). Il n'y a alors aucun avantage à utiliser ces techniques de préconditionnement.

Enfin, on montre que l'utilisation systématique de la méthode FCP pour initialiser le 1er jeu de composants avant l'algorithme de Newton permet un gain significatif en convergence, surtout lorsque le conditionnement de la matrice $\tilde{\mathbf{Z}}^n$ est grand.

2.3.4. PRISE EN COMPTE DE LA VARIATION DES COEFFICIENTS D'ACTIVITE

Dans le calcul usuel de la matrice jacobienne $\mathbf{Z}$, tel que présenté dans l'équation (2.17), il est supposé que les coefficients d'activités $ln(\gamma_i)$ des espèces dissoutes sont quasiment constant pour les variations des activités de composants envisagées. Or dans certains cas, cette hypothèse n'est pas valide, et nous avons tenté de développer une approche prenant en compte ces éléments. Si l'on reprend la définition de la jacobienne en incluant la variation des coefficients d'activité, on obtient la relation suivante :

$$\tilde{Z}^n_{j,k} = \frac{\partial}{\partial \xi_k}\left\{\sum_{i=1}^{Nc} b_{i,j}\cdot exp\left(ln(K_i)+\sum_{h=1}^{Nx} a_{i,h}\cdot \xi_h^n - ln(\gamma_i^n)\right)\right\} \quad \forall j,k=(1,...Nx)^2 \tag{2.49}$$

Soit

$$\tilde{Z}^n_{j,k} = \sum_{i=1}^{Nc} b_{i,j}\cdot C_i^n \cdot \left(a_{i,k} - \frac{\partial\left[ln(\gamma_i^n)\right]}{\partial \xi_k}\right) \quad \forall j,k=(1,...Nx)^2 \tag{2.50}$$

Cette relation est obtenue à l'aide de l'élément suivant :

$$\frac{\partial C_i^n}{\partial \xi_k} = \frac{\partial}{\partial \xi_k}\left\{exp\left(ln(K_i)+\sum_{h=1}^{Nx} a_{i,h}\cdot \xi_h^n - ln(\gamma_i^n)\right)\right\} = C_i^n\cdot\left(a_{i,k} - \frac{\partial\left[ln(\gamma_i^n)\right]}{\partial \xi_k}\right) \tag{2.51}$$

Dans le cas d'un modèle de Davis de correction d'activité, le coefficient γ_i est donné par (1.99)

$$ln(\gamma_i) = -ln(10)Az_i^2\left(\frac{\sqrt{I}}{1+\sqrt{I}} - BI\right) \tag{1.99}$$

Et la force ionique I par (1.98), où NcA est le nombre d'espèces chimique en phase aqueuse :

$$I = \frac{1}{2}\sum_{i=1}^{NcA} z_i^2\cdot[C_i] \tag{1.98}$$

On a alors :

$$\frac{\partial\left[ln\left(\gamma_i^n\right)\right]}{\partial\xi_k}=\frac{\partial\left[ln\left(\gamma_i^n\right)\right]}{\partial I^n}\frac{\partial I^n}{\partial\xi_k} \tag{2.52}$$

Le terme de dérivation de la force ionique donne, en utilisant la relation (2.51) :

$$\frac{\partial I^n}{\partial\xi_k}=\frac{1}{2}\sum_{i=1}^{NcA}z_i^2\cdot\frac{\partial\left[C_i^n\right]}{\partial\xi_k}=\frac{1}{2}\sum_{i=1}^{NcA}z_i^2\cdot C_i^n\cdot\left(a_{i,k}-\frac{\partial\left[ln\left(\gamma_i^n\right)\right]}{\partial\xi_k}\right) \tag{2.53}$$

Le terme de dérivation du coefficient d'activité selon la force ionique donne :

$$\frac{\partial\left[ln\left(\gamma_i^n\right)\right]}{\partial I^n}=-ln(10)Az_i^2\frac{\partial}{\partial I^n}\left(\frac{\sqrt{I^n}}{1+\sqrt{I^n}}-BI^n\right)=-\frac{1}{2}ln(10)Az_i^2\left[\frac{1}{\sqrt{I^n}}\frac{1}{\left(1+\sqrt{I^n}\right)^2}-B\right] \tag{2.54}$$

On obtient alors pour la dérivée du coefficient d'activité par l'activité du composant :

$$\frac{\partial\left[ln\left(\gamma_i^n\right)\right]}{\partial\xi_k}=-\frac{1}{2}ln(10)Az_i^2\left[\frac{1}{\sqrt{I^n}}\frac{1}{\left(1+\sqrt{I^n}\right)^2}-B\right]\cdot\frac{1}{2}\sum_{p=1}^{NcA}z_p^2\cdot C_p^n\cdot\left(a_{p,k}-\frac{\partial\left[ln\left(\gamma_p^n\right)\right]}{\partial\xi_k}\right) \tag{2.55}$$

Ce terme peut être séparé en 2 partie, l'une explicite (ne dépendant pas de $ln\left(\gamma_p^n\right)$), l'autre implicite :

$$\frac{\partial\left[ln\left(\gamma_i^n\right)\right]}{\partial\xi_k}=N_i+\sum_{p=1}^{NcA}M_{i,p}\cdot\frac{\partial\left[ln\left(\gamma_p^n\right)\right]}{\partial\xi_k} \tag{2.56}$$

Avec

$$N_i=-\frac{1}{4}ln(10)Az_i^2\left[\frac{1}{\sqrt{I^n}}\frac{1}{\left(1+\sqrt{I^n}\right)^2}-B\right]\cdot\sum_{p=1}^{NcA}z_p^2\cdot C_p^n\cdot a_{p,k} \tag{2.57}$$

Et

$$M_{i,p}=\frac{1}{4}ln(10)Az_i^2\left[\frac{1}{\sqrt{I^n}}\frac{1}{\left(1+\sqrt{I^n}\right)^2}-B\right]\cdot\sum_{p=1}^{NcA}z_p^2\cdot C_p^n \tag{2.58}$$

On a alors une formulation matricielle de la relation (2.55) :

$$\frac{\partial}{\partial\xi_k}\left[ln\left(\gamma^n\right)\right]=\mathbf{N}+\mathbf{M}\cdot\frac{\partial}{\partial\xi_k}\left[ln\left(\gamma^n\right)\right] \tag{2.59}$$

Il est alors possible d'obtenir le terme de dérivée du coefficient d'activité par résolution du système :

$$\frac{\partial}{\partial\xi_k}\left[ln\left(\gamma^n\right)\right]=\left(\mathbf{Id}-\mathbf{M}\right)^{-1}\cdot\mathbf{N} \tag{2.60}$$

Ainsi, pour calculer les termes de la matrice $\mathbf{Z}$, qui est de dimension $Nx \times Nx$, il devient nécessaire d'inverser le système (2.60) qui est de dimension $NcA \times NcA$. Il convient de rappeler que le nombre de composants Nx est usuellement assez petit, 10 à 50 éléments, alors que le nombre d'espèces dissoutes est souvent beaucoup plus grand, parfois plusieurs centaines.

Les résultats préliminaires obtenus en incluant ces dérivées des coefficients d'activité montrent un très léger gain en termes de nombre d'itération pour certains cas, mais en termes de temps de calcul, la résolution du système (2.60) est trop importante pour que le bénéfice de cette approche soit sensible. Ce travail nécessite encore une étude plus systématique, notamment le test de nombreux cas chimiques différents.

2.3.5. RECHERCHE D'UNE BASE DE COMPOSANTS OPTIMALE

Suite à ces premiers travaux, il nous semblait clair que la source des difficultés numériques est liée à un mauvais conditionnement de la matrice jacobienne. Ce mauvais conditionnement entrainerait l'échec de la procédure d'inversion, la divergence de la procédure de recherche vers des valeurs non physiques. Une première piste de recherche a été suivie lors du stage de master de Thierry Schwoertzig : nous avons envisagé la possibilité que toutes les bases de composant possible pour décrire un sytème chimique donné ne soient pas équivalentes. L'hypothèse était que certaines bases conduiraient à des matrices jacobiennes mieux conditionnées que d'autres.

En nous basant sur l'analogie très forte entre le concept mathématique d'espace vectoriel (et des bases de vecteurs) et la formulation des systèmes chimiques sous forme d'espèces et de composants, les outils numériques de passage d'une base de composant à une autre ont été développés.

2.3.5.1. Développement mathématiques

Les éléments présentés ici ont été proposés par Westall [91, 180] dans le cadre du changement d'un seul composant à l'intérieur de la base. Nous présentons ici un formalisme plus général autorisant de modifier plusieurs composants.

Si l'on considère un système chimique exprimé dans une base de composants $\mathbf{X}_{old}$, il est nécessaire d'avoir une matrice stœchiométrique, un vecteur de constantes d'équilibre, un vecteur de concentrations totales en composant et un vecteur de composant. S'il est possible de construire une autre base de composants, $\mathbf{X}_{new}$, pour ce système chimique, alors il existe une matrice de passage $\mathbf{P}_B$ entre les bases $\mathbf{X}_{old}$ et $\mathbf{X}_{new}$.

$$\mathbf{X}_{new} = \mathbf{P}_B \cdot \mathbf{X}_{old} \tag{2.61}$$

L'équation (2.61) représente les réactions chimiques décrivant la formation des composants de la nouvelle base à partir des composants de l'ancienne base. La matrice de passage $\mathbf{P}_B$ est donc composée des coefficients stœchiométriques relatifs à la conservation de la matière. D'autre part, ces réactions sont décrites par des lois d'action de masse, que l'on peut écrire sous forme logarithmique :

$$\boldsymbol{\xi}_{new} = ln\left(\tilde{\mathbf{K}}_{old}\right) + \mathbf{P}_A \cdot \boldsymbol{\xi}_{old} \tag{2.62}$$

où $\mathbf{P}_A$ est la matrice de passage entre les deux bases composée des coefficients stœchiométriques relatifs à la loi d'action de masse et $\tilde{\mathbf{K}}_{old}$ est la restriction du vecteur des constantes d'équilibre aux seuls composants $\mathbf{X}_{new}$.

On peut alors reconstruire une matrice stœchiométrique, ainsi que les vecteurs de concentrations totales et de constantes d'équilibre.

Nouvelles matrices stœchiométriques

Les réactions chimiques de formation des espèces en fonction du jeu de composant $\mathbf{X}_{old}$ s'écrivent sous forme matricielle :

$$\mathbf{C} = \mathbf{B}_{old} \cdot \mathbf{X}_{old} \tag{2.63}$$

En utilisant l'équation (2.61) et le fait que les matrices de passage sont inversibles, nous pouvons exprimer ces réactions de formation des espèces en fonction du jeu de composant $\mathbf{X}_{new}$:

$$\mathbf{C} = \mathbf{B}_{old} \cdot \mathbf{P}_B^{-1} \cdot \mathbf{X}_{new} \tag{2.64}$$

Les matrices stœchiométriques relatives aux lois de conservation et aux lois d'action de masse s'expriment alors dans la base de composants $\mathbf{X}_{new}$:

$$\mathbf{B}_{new} = \mathbf{B}_{old} \cdot \mathbf{P}_B^{-1} \text{ et } \mathbf{A}_{new} = \mathbf{A}_{old} \cdot \mathbf{P}_A^{-1} \tag{2.65}$$

Nouveaux vecteurs de concentrations totales

Exprimées dans la bases de composants $\mathbf{X}_{new}$, les lois de conservations des concentrations totales en composants s'écrivent :

$$\mathbf{T}_{new} = \mathbf{B}_{new}^{\ T} \cdot \mathbf{C} \tag{2.66}$$

En faisant apparaitre la matrice stœchiométrique $\mathbf{B}_{old}$ à l'aide de l'équation (2.65), on obtient :

$$\mathbf{T}_{new} = \left(\mathbf{B}_{old} \cdot \mathbf{P}_B^{-1}\right)^T \cdot \mathbf{C} = \left(\mathbf{P}_B^{-1}\right)^T \cdot \mathbf{B}_{old}^{\ T} \cdot \mathbf{C} \tag{2.67}$$

Ce qui permet d'exprimer la concentration totale dans la base $\mathbf{X}_{new}$ en fonction de celle dans la base $\mathbf{X}_{old}$:

$$\mathbf{T}_{new} = \left(\mathbf{P}_B^{-1}\right)^T \cdot \mathbf{T}_{old} \tag{2.68}$$

Nouveaux vecteurs de constantes d'équilibre

Exprimées sous forme logarithmique dans la base de composants $\mathbf{X}_{new}$, les loi d'actions de masse décrivant la formation des espèces s'écrivent :

$$ln\left(\mathbf{C}\right) = ln\left(\mathbf{K}_{new}\right) + \mathbf{A}_{new} \cdot \boldsymbol{\xi}_{new} \tag{2.69}$$

Les relations (2.62) et (2.65) permettent de faire apparaitre les éléments relatifs à la base $\mathbf{X}_{old}$ dans cette loi d'action de masse :

$$ln(\mathbf{C}) = ln(\mathbf{K}_{new}) + \mathbf{A}_{old} \cdot \mathbf{P}_A^{-1} \cdot \left[ln(\tilde{\mathbf{K}}_{old}) + \mathbf{P}_A \cdot \boldsymbol{\xi}_{old} \right] \quad (2.70)$$

Par simplifications, on obtient :

$$ln(\mathbf{C}) = ln(\mathbf{K}_{new}) + \mathbf{A}_{old} \cdot \mathbf{P}_A^{-1} \cdot ln(\tilde{\mathbf{K}}_{old}) + \mathbf{A}_{old} \cdot \mathbf{P}_A^{-1} \cdot \mathbf{P}_A \cdot \boldsymbol{\xi}_{old} \quad (2.71)$$

Soit :

$$ln(\mathbf{C}) = ln(\mathbf{K}_{new}) + \mathbf{A}_{old} \cdot \mathbf{P}_A^{-1} \cdot ln(\tilde{\mathbf{K}}_{old}) + \mathbf{A}_{old} \cdot \boldsymbol{\xi}_{old} = ln(\mathbf{K}_{old}) + \mathbf{A}_{old} \cdot \boldsymbol{\xi}_{old} \quad (2.72)$$

Par identifications, les constantes d'équilibre dans la base $\mathbf{X}_{new}$ se déterminent de la manière suivante :

$$ln(\mathbf{K}_{new}) = ln(\mathbf{K}_{old}) - \mathbf{A}_{old} \cdot \mathbf{P}_A^{-1} \cdot ln(\tilde{\mathbf{K}}_{old}) \quad (2.73)$$

2.3.5.2. Applications et perspectives

Sur la base de ces outils numériques, plusieurs systèmes chimiques ont été testés, et pour chacun d'eux nous avons comparé différentes bases de composants. Les matrices jacobiennes ont été comparées ainsi que la vitesse de convergence de l'algorithme. L'hypothèse initiale que certaines bases de composants conduisaient à une résolution plus facile que d'autres a bien été confirmée. Cependant nous n'avons pas réussi à dégager de critère permettant de sélectionner à priori une base performante.

Plus récemment, ce concept de changement de base dans l'espace des composants a été appliqué avec succès Hoffmann *et al.* [105]. Au lieu de chercher une base permettant une matrice jacobienne mieux conditionnée (aspect moins critique dans le cadre de l'approche globale choisie par cette équipe), ils ont cherché des bases de composants induisant un découplage entre les équations de transport de certains composants.

Cette approche peut nous être utile pour trois raisons :

(i) Dans le cadre de notre recherche d'une base de composants générant des matrices stœchiométriques mieux conditionnées, il faut retenir l'ouverture offerte par les travaux d'Hoffmann et al. : il n'est pas nécessaire que les composants choisis aient une existence physique. L'exploration d'un ensemble plus vaste de base de composants pourrait nous permettre de trouver un critère à priori de sélection des bases.

(ii) La démarche de découplage des équations de transport réactif développée par Hoffmann *et al.* inclus une analyse à priori des bases de composants pour déterminer laquelle conduit au système le plus découplé. Est-il possible de s'inspirer de cette démarche pour développer un critère de sélection des bases ?

(iii) En élargissant l'analyse menée dans ce chapitre et en revenant à l'objectif global de transport réactif, rien n'empêche de formuler une démarche de résolution par séparation d'opérateurs sur une base de composant choisie selon les critères de découplage proposés par Hoffmann *et al.* De ce fait, les erreurs de séparation d'opérateurs seraient localisées sur les seuls composants couplés. La question induite est évidement de savoir si ces erreurs sont alors plus faibles ou plus fortes que lorsqu'elles affectent davantage de composants. Dans le cadre d'une résolution par séparation d'opérateur itérative, il ne serait plus nécessaire de calculer les itérations de transport sur les composants découplés. La question

induite est alors de savoir si le temps de calcul global est alors réduit par rapport à une résolution sur une autre base de composants.

Enfin, les travaux menés dans le cadre du benchmark de MoMas ont amené une nouvelle formulation de l'opérateur de chimie (voir § 2.3.3.1). Avec cette nouvelle formulation, la matrice jacobienne est calculée différemment. Il est possible que cette nouvelle formulation soit plus facile à étudier et que nous puissions alors trouver un critère de sélection des bases à priori.

2.4. RESOLUTION DE SYSTEMES COUPLES CINETIQUE-EQUILIBRE

La description des phénomènes chimiques à l'équilibre thermodynamique instantané est une approximation (Local Equilibrium Approximation : LEA) qui permet de s'affranchir d'une description cinétique des réactions. Dans le cadre de notre étude, la modélisation numérique des transferts réactifs en milieu poreux, une telle approximation permet une diminution importante de la complexité. Comme cela a été présenté (§ 1.4.3), on peut alors réduire le nombre d'équations de transport à résoudre en ne transportant que les composants.

Cependant, cette approximation n'est possible que lorsque la vitesse des réactions est grande devant celle du transport. Pour de nombreuses réactions, cela est tout à fait valable, même pour des écoulements assez rapides. En effet, les temps caractéristiques des réactions acide-bases par exemple sont de l'ordre de 10^{-9} à 10^{-6} s.

En général, on utilise les nombres de Damköhler (adimentionnels) pour comparer les temps caractéristiques du transport et de réaction. Le nombre de Damköhler Da_I représente le rapport entre la vitesse d'une réaction et le flux de matière entrant dans un réacteur :

$$Da_I = \frac{k \cdot [c]^n \cdot V}{\Phi} \tag{2.74}$$

où k est la constante cinétique de la réaction, n sont ordre, V le volume du réacteur et Φ le flux de matière entrant dans le réacteur. Si l'on applique cette définition, à une réaction d'ordre 1 ayant lieu dans un milieu poreux de longueur L et de porosité ω avec une vitesse de Darcy U, on obtient

$$Da_I = \frac{k \cdot [c] \cdot L \cdot S}{\omega \cdot U \cdot [c] \cdot S} = \frac{k \cdot L}{\omega \cdot U} \tag{2.75}$$

Pour d'autres réactions, il est cependant nécessaire de prendre en compte l'aspect cinétique du problème. Il s'agit entre autre des phénomènes biologiques (croissance microbienne ou biodégradation), de nombreux phénomènes aux interfaces (dissolution d'un gaz, dissolution d'un minéral ou précipitation d'une phase secondaire). Certains auteurs [104] ont choisi de décrire l'intégralité des phénomènes sous forme cinétique. Cela se traduit par un nombre d'équations de transport égal au nombre d'espèces mobiles et par des pas de temps assez petits car la résolution d'un système différentiel et algébrique est limitée par la vitesse de réaction la plus rapide. Ce point peut expliquer le temps de calcul important nécessaire au code MIN3P pour résoudre le test du benchmark MoMaS [84]. En plus de ces considérations numériques, l'hypothèse de l'équilibre local est souvent utilisée pour des raisons pratiques : les bases de données sont beaucoup plus riches et accessibles pour les phénomènes chimiques à l'équilibre que pour la description cinétique [209].

Ainsi, il nous semble important de pouvoir associer, au sein d'un même modèle, une description cinétique de certains phénomènes chimiques et une description à l'équilibre instantané pour d'autres. Il faut également que les vitesses de réactions puissent faire appel à la concentration de certaines espèces chimiques décrites par LEA : le proton H^+ en est l'exemple le plus évident.

2.4.1. CRITERES DE CHOIX

2.4.1.1. Test expérimental

Plusieurs auteurs se sont penchés sur la problématique de la validité ou non de l'approximation de l'équilibre local. Expérimentalement, on peut considérer que la formulation LEA est valide si une modification de la vitesse d'écoulement ne modifie pas la forme du profil de concentration ou de la courbe d'élution. Dans ce cas, ces courbes tracées en temps ou en coordonnées réduites se superposent.

Un exemple en est donné par M. Bueno *et al.* [95], qui étudient la sorption du tributhyl étain (TBT) sur un sable de quartz. Ces auteurs montrent que les courbes d'élution du TBT se superposent (en coordonnées réduites) pour différentes vitesses de pores allant de 0,05 cm/min jusqu'à 5 cm/min (voir Figure 2-2). On en conclut que, dans ce cas, les vitesses de réactions sont très rapides par rapport aux changements de concentrations induits par l'écoulement.

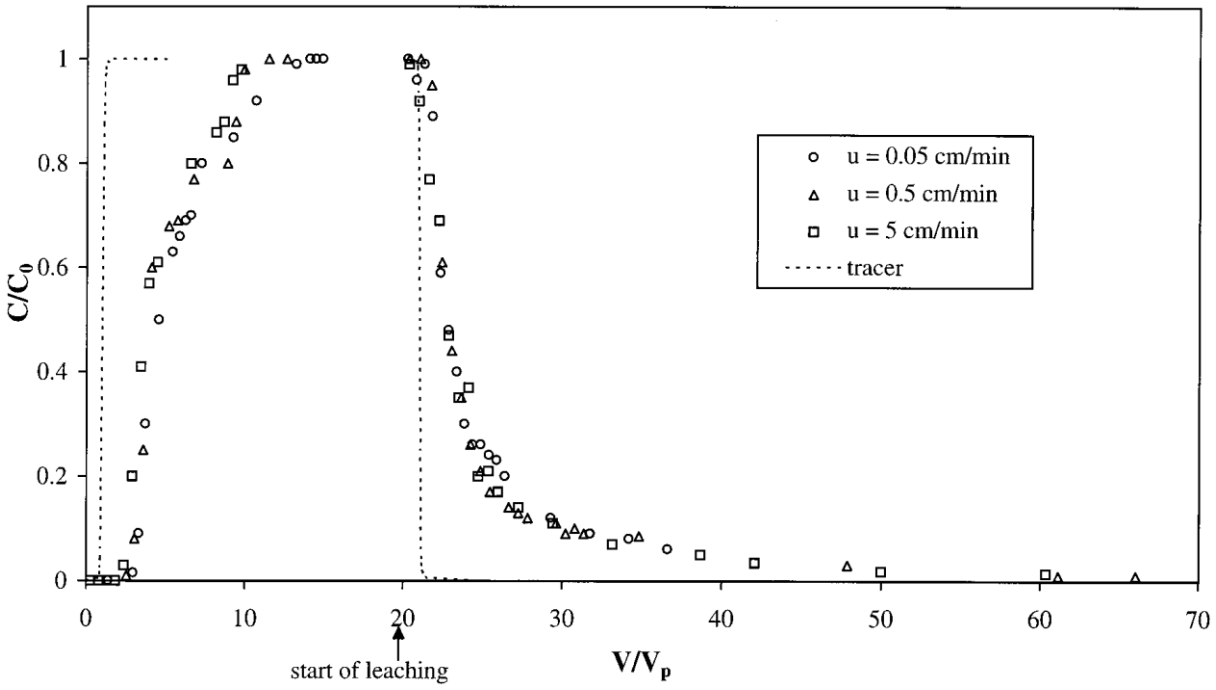


Figure 2-2 : Courbes d'élution du TBT
à travers une colonne de sable de quartz en fonction de la vitesse de pore
V volume injecté, V_p volume de pore de la colonne ; C concentration en TBT, concentration en entrée ; u vitesse de pore. (source : Bueno *et al.* [95])

2.4.1.2. Test numérique

Numériquement, certains auteurs ont étudié l'influence d'une formulation LEA et d'une formulation cinétique d'un même problème. Par exemple, Jennings et Kirkner [210] ont simulé des courbes d'élution et des profils de concentration pour un système chimique à plusieurs composants incluant de la complexation en phase aqueuse, de la sorption avec une compétition entre deux composants ou de l'échange d'ions. En comparant les courbes d'élution ou les profils de concentration obtenus par une formulation cinétique avec ceux obtenus sous LEA, ils montrent que les deux formulations deviennent équivalentes si le nombre de Damköhler devient suffisamment grand. Ces auteurs proposent une valeur de nombre de Damköhler de 100 comme limite inférieure de validité de l'approximation de l'équilibre instantané comme cela peut se

voir sur la Figure 2-3. Cependant, ils soulignent que cette valeur est très dépendante du type de réaction, des conditions initiales et aux limites etc.

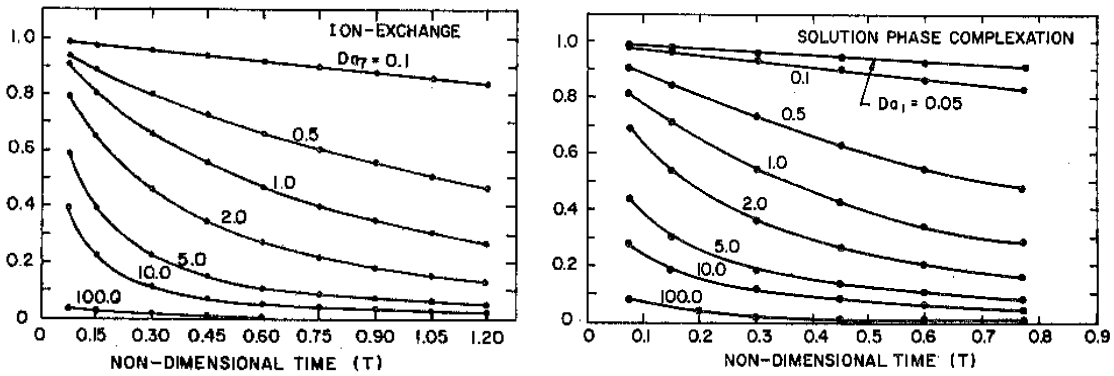


Figure 2-3 : Différence par rapport à une solution par équilibre instantané en fonction du temps caractéristique et pour différentes valeurs de nombre de Damköhler.
A gauche pour une réaction d'échange d'ion ; à droite pour une réaction de sorption.
(source : Jennings et Kirkner [210])

2.4.1.3. Etudes théoriques

Rubin [211] étudie les différents types de réactions possibles et propose une classification ainsi que des modèles pour décrire chacun des types décrits. La classification proposée repose fortement sur la validité ou non de l'hypothèse LEA. En effet, comme on peut le voir sur la Figure 2-4, deux grandes classes se dessinent, l'une avec des réactions réversibles et suffisamment rapides (sous-entendu pour que l'hypothèse LEA soit valide) et l'autre avec des réactions irréversibles et/ou pas assez rapides.

Valocchi [212] étudie l'influence d'une formulation cinétique ou LEA en comparant les moments temporels des courbes d'élutions dans le cas d'un transport mono dimensionnel en milieu homogène. Il montre que la limite de validité de l'hypothèse LEA dépend de nombreux facteurs. Certes les nombres de Damköhler font partie des facteurs importants identifiés dans ce travail, mais le nombre de Péclet et le coefficient de retard jouent également un rôle important.

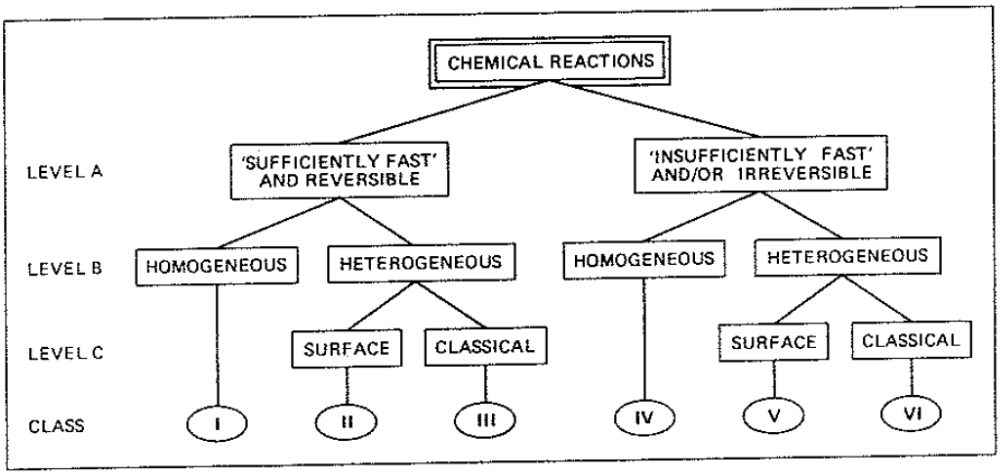


Figure 2-4 : Classification des différentes réactions chimiques selon Rubin [211]

Bahr et Rubin [213] proposent une approche du problème basée sur la séparation du terme influencé par la cinétique (Separation of the Kinetically Influenced Term : SKIT). Cette démarche est suffisamment générale pour que les auteurs l'appliquent à de nombreux phénomènes chimiques : réaction d'ordre 1, cinétique de Langmuir, cinétique d'ordre n, diffusion et isotherme linéaire ou sorption multi-composant… Nous présenterons ici cette démarche dans le cadre d'une cinétique d'ordre 1.

La première étape nécessite l'écriture des équations de transport réactif pour l'approche cinétique et pour la formulation LEA. Si l'on considère une réaction d'immobilisation d'une espèce chimique C pour donner une espèce fixée S :

$$C \underset{k_r}{\overset{k_f}{\rightleftharpoons}} S \tag{2.76}$$

La vitesse de cette réaction d'ordre 1 s'écrit :

$$r = \omega \frac{d[c]}{dt} = -\rho_s \frac{d[s]}{dt} = -k_f [c] + k_r [s] \tag{2.77}$$

Dans le cas d'une formulation à l'équilibre, la vitesse de cette réaction est nulle et donc :

$$k_f [c] = k_r [s] \text{ soit } [s] = \frac{k_f}{k_r}[c] = K[c] \tag{2.78}$$

où K est la constante d'équilibre de cette réaction.

Dans le cas d'une formulation cinétique de ce problème de transport réactif en monodimensionnel, l'équation de conservation donne :

$$\frac{\partial [c]}{\partial t} = D \frac{\partial^2 [c]}{\partial x^2} - U \frac{\partial [c]}{\partial x} - \frac{k_f}{\omega}[c] + \frac{k_r}{\omega}[s] \tag{2.79}$$

et

$$\rho_s \frac{\partial [s]}{\partial t} = k_f [c] - k_r [s] \tag{2.80}$$

Si l'on divise l'équation (2.80) par la porosité et que l'on ajoute cette relation à l'équation (2.79) on obtient la formulation LEA :

$$\frac{\partial [c]}{\partial t} + \frac{\rho_s}{\omega} \frac{\partial [s]}{\partial t} = D \frac{\partial^2 [c]}{\partial x^2} - U \frac{\partial [c]}{\partial x} \tag{2.81}$$

Lors de la seconde étape, nous allons éliminer la concentration en espèce fixée $[s]$ dans chacune des équations (2.79) et (2.81).

En utilisant la loi d'action de masse (2.78), on peut éliminer $[s]$ de la relation (2.81) pour faire apparaitre le facteur de retard $R = 1 + K \frac{\rho_s}{\omega}$:

$$\left(1 + K \frac{\rho_s}{\omega}\right) \frac{\partial [c]}{\partial t} = D \frac{\partial^2 [c]}{\partial x^2} - U \frac{\partial [c]}{\partial x} \tag{2.82}$$

On obtient alors l'équation retardée dans la formulation LEA :

$$\frac{\partial[c]}{\partial t} = \frac{D}{R}\frac{\partial^2[c]}{\partial x^2} - \frac{U}{R}\frac{\partial[c]}{\partial x} \tag{2.83}$$

En isolant $[s]$ dans l'équation (2.79), on obtient une expression explicite de $[s]$:

$$[s] = \frac{\omega}{k_r}\left(\frac{\partial[c]}{\partial t} - D\frac{\partial^2[c]}{\partial x^2} + U\frac{\partial[c]}{\partial x}\right) + \frac{k_f}{k_r}[c] \tag{2.84}$$

La troisième étape consiste en la dérivation temporelle de la relation (2.84) :

$$\frac{\partial[s]}{\partial t} = \frac{\omega}{k_r}\frac{\partial}{\partial t}\left(\frac{\partial[c]}{\partial t} - D\frac{\partial^2[c]}{\partial x^2} + U\frac{\partial[c]}{\partial x}\right) + \frac{k_f}{k_r}\frac{\partial[c]}{\partial t} \tag{2.85}$$

Lors de la quatrième étape, on substitue la relation (2.85) dans la relation (2.81) et l'on transforme l'équation obtenue pour faire apparaître l'équation LEA retardée (2.83) et un terme correctif.

La substitution de (2.85) dans (2.81) donne :

$$\frac{\partial[c]}{\partial t} + \frac{\rho_s}{\omega}\left\{\frac{\omega}{k_r}\frac{\partial}{\partial t}\left(\frac{\partial[c]}{\partial t} - D\frac{\partial^2[c]}{\partial x^2} + U\frac{\partial[c]}{\partial x}\right) + \frac{k_f}{k_r}\frac{\partial[c]}{\partial t}\right\} = D\frac{\partial^2[c]}{\partial x^2} - U\frac{\partial[c]}{\partial x} \tag{2.86}$$

Soit

$$\frac{\partial[c]}{\partial t} + \frac{\rho_s}{\omega}\frac{k_f}{k_r}\frac{\partial[c]}{\partial t} = D\frac{\partial^2[c]}{\partial x^2} - U\frac{\partial[c]}{\partial x} - \frac{\rho_s}{k_r}\left\{\frac{\partial}{\partial t}\left(\frac{\partial[c]}{\partial t} - D\frac{\partial^2[c]}{\partial x^2} + U\frac{\partial[c]}{\partial x}\right)\right\} \tag{2.87}$$

Et donc

$$R\frac{\partial[c]}{\partial t} = D\frac{\partial^2[c]}{\partial x^2} - U\frac{\partial[c]}{\partial x} - \frac{\rho_s}{k_r}\left\{\frac{\partial}{\partial t}\left(\frac{\partial[c]}{\partial t} - D\frac{\partial^2[c]}{\partial x^2} + U\frac{\partial[c]}{\partial x}\right)\right\} \text{ car } R = 1 + K\frac{\rho_s}{\omega} = 1 + \frac{\rho_s}{\omega}\frac{k_f}{k_r} \tag{2.88}$$

On fait alors apparaitre l'équation retardée :

$$\frac{\partial[c]}{\partial t} = \frac{D}{R}\frac{\partial^2[c]}{\partial x^2} - \frac{U}{R}\frac{\partial[c]}{\partial x} - \frac{\rho_s}{k_r}\cdot\frac{1}{1 + \frac{\rho_s}{\omega}\frac{k_f}{k_r}}\left\{\frac{\partial}{\partial t}\left(\frac{\partial[c]}{\partial t} - D\frac{\partial^2[c]}{\partial x^2} + U\frac{\partial[c]}{\partial x}\right)\right\} \tag{2.89}$$

Par réarrangement, on obtient la formulation SKIT :

$$\frac{\partial[c]}{\partial t} = \frac{D}{R}\frac{\partial^2[c]}{\partial x^2} - \frac{U}{R}\frac{\partial[c]}{\partial x} - \frac{1}{\frac{k_r}{\rho_s} + \frac{k_f}{\omega}}\left\{\frac{\partial}{\partial t}\left(\frac{\partial[c]}{\partial t} - D\frac{\partial^2[c]}{\partial x^2} + U\frac{\partial[c]}{\partial x}\right)\right\} \tag{2.90}$$

La différence entre la formulation LEA (2.83) et la formulation cinétique SKIT est constituée par le terme additionnel

$$\frac{1}{\frac{k_r}{\rho_s}+\frac{k_f}{\omega}}\left\{\frac{\partial}{\partial t}\left(\frac{\partial[c]}{\partial t}-D\frac{\partial^2[c]}{\partial x^2}+U\frac{\partial[c]}{\partial x}\right)\right\} \tag{2.91}$$

Les formulations LEA et SKIT peuvent être présentées sous formes adimensionnelles en introduisant les termes suivants :

Longueur adimensionnelle :

$$X=\frac{x}{L} \tag{2.92}$$

où L est une longueur de référence, souvent la longueur du milieu (colonne de laboratoire).

Temps réduit :

$$T=\frac{U}{R}\frac{t}{L} \tag{2.93}$$

Concentration réduite :

$$C=\frac{[c]}{[c_0]} \tag{2.94}$$

où $[c_0]$ est une concentration de référence, souvent la concentration injectée à la limite du domaine.

Le nombre de Péclet :

$$Pe=\frac{UL}{D} \tag{2.95}$$

Un nombre de Damköhler I pour la sorption :

$$Da_f=\frac{k_f L}{\omega U} \tag{2.96}$$

Un nombre de Damköhler I pour la désorption :

$$Da_r=\frac{k_r L}{\rho_s U} \tag{2.97}$$

En variables adimensionnelles, la formulation LEA s'écrit :

$$\frac{\partial C}{\partial T}-\frac{1}{Pe}\frac{\partial^2 C}{\partial X^2}+\frac{\partial C}{\partial X}=0 \tag{2.98}$$

La formulation SKIT adimensionnelle s'écrit :

$$\frac{\partial C}{\partial T} - \frac{1}{Pe}\frac{\partial^2 C}{\partial X^2} + \frac{\partial C}{\partial X} = -\frac{1}{Da_f + Da_r}\left\{\frac{\partial}{\partial t}\left(\frac{1}{1+\frac{Da_f}{Da_r}}\frac{\partial C}{\partial T} - \frac{1}{Pe}\frac{\partial^2 C}{\partial X^2} + \frac{\partial C}{\partial X}\right)\right\} \quad (2.99)$$

Bahr et Rubin [213] montrent ainsi que la différence entre une formulation LEA et une formulation SKIT est contenue dans le second membre composé d'un préfixe et de la dérivée temporelle. On retrouve ainsi certaines considérations intuitives : lorsque le régime permanent est atteint, il n'y a pas de différence entre une formulation cinétique et une formulation à l'équilibre. Ils montrent également que la différence entre les deux formulations varie dans l'espace et le temps et dépend de tous les paramètres physico-chimiques comme l'avait mis en évidence Valocchi [212]. Cependant ces auteurs montrent également que la somme des nombres de Damköhler $Da_f + Da_r$ est plutôt prédominant dans la validité ou non de l'hypothèse LEA et ils confirment la valeur limite de 100 proposée par Jennings et Kirkner [210].

La question serait alors de pouvoir généraliser cette approche pour pouvoir proposer une classification *automatique* des réactions conduisant, au fil de la simulation, à la prise en compte ou au rejet du *caractère cinétique* de telle ou telle réaction.

2.4.2. OUTILS NUMERIQUES

Dans le cadre de l'approche choisie ici, une approche par séparation d'opérateurs (voir 2.2) les termes de transport n'apparaissent dans l'opérateur de chimie cinétique que comme un terme puit-source. Le verrou se situe donc dans la façon d'associer la description à l'équilibre et celle cinétique. De très nombreux travaux existent sur les différentes méthodes pour calculer les systèmes chimiques cinétiques, notamment en modélisation de la chimie atmosphérique, où la rapidité de l'évolution des conditions de d'écoulement, de température, de pression et d'ensoleillement impose une description cinétique de tous les phénomènes chimiques. A cause d'une très grande différence de temps caractéristiques entre les différents phénomènes chimiques, on parle alors d'un système d'équations différentielles ordinaires raides. Plusieurs études sur la chimie atmosphérique [214-216] ont montré que les solveurs implicites sont beaucoup plus performants que les solveurs explicites sur ces systèmes.

Malheureusement, la contrainte qui nous est imposée d'associer équilibre et cinétique rend la mise en place de méthodes implicites extrêmement complexe. Nous n'avons à ce jour pas encore trouvé ni d'application opérationnelle d'un solveur implicite, ni même de méthode numérique permettant une résolution implicite d'un système composé d'équations différentielles ordinaires raides et d'équations algébriques fortement non linéaires. Nous avons donc choisi d'orienter nos efforts vers les méthodes d'adaptation du pas de temps.

2.4.2.1. Extrapolation de Richardson

Le principe de base de cette méthode est de résoudre un système différentiel en 2 étapes : une première résolution sur un pas de temps complet, et une seconde résolution en 2 demi pas de temps. Ainsi, il est nécessaire, à chaque pas de temps, de résoudre 3 fois le système. Cependant, nous verrons que cet investissement peut se révéler extrêmement avantageux. Cette méthode est proposée par Richardson [217, 218] et est détaillée par Hairer *et al.* [219].

Cette méthode a été adaptée à la résolution des systèmes différentiels décrivant du transport réactif à l'équilibre instantané en milieu poreux saturé et à des systèmes décrivant le transfert de masse en milieu poreux non saturé. Le détail de ces adaptations est donné en Annexe 6. Voici le principe général de cette méthode :

Soit l'équation (2.100), représentant la forme générale d'une ODE, d'une PDE, ou d'un système de cette nature :

$$\frac{dc}{dt} = f(c) \tag{2.100}$$

L'équation (2.100) peut être discrétisée en temps (et en espace si nécessaire) pour donner la relation suivante entre les pas de temps n et $n+1$ de durée Δt :

$$\frac{c^{n+1} - c^n}{\Delta t} = f(c) \tag{2.101}$$

2.4.2.1.1. Extrapolation

L'équation (2.101) étant résolue à l'aide d'une méthode numérique d'ordre p en temps, la différence entre la solution exacte et la solution calculée au pas de temps $n+1$ est donnée par :

$$c_{Ex}^{n+1} - c^{n+1,*} = A \cdot \Delta t^{p+1} + \mathcal{O}\left(\Delta t^{p+2}\right) \tag{2.102}$$

si l'équation est résolue en un seul pas de temps. Si l'équation (2.101) est résolue en 2 demi pas de temps, la différence entre solution exacte et solution approchée au demi pas de temps est donnée par :

$$c_{Ex}^{n+1/2} - c^{n+1/2} = A \cdot \left(\frac{\Delta t}{2}\right)^{p+1} + \mathcal{O}\left(\Delta t^{p+2}\right) \tag{2.103}$$

Au second demi pas de temps, l'erreur obtenue est la somme de l'erreur transportée depuis le 1er demi pas de temps et de celle provenant du second :

$$c_{Ex}^{n+1} - c^{n+1,**} = \left[I + \frac{\partial f}{\partial c}\Delta t + \mathcal{O}\left(\Delta t^2\right)\right] \cdot \left[A \cdot \left(\frac{\Delta t}{2}\right)^{p+1} + \mathcal{O}\left(\Delta t^{p+2}\right)\right] + A \cdot \left(\frac{\Delta t}{2}\right)^{p+1} + \mathcal{O}\left(\Delta t^{p+2}\right) \tag{2.104}$$

En simplifiant l'équation (2.104) et en éliminant les termes d'ordre supérieurs à $p+2$, on obtient :

$$c_{Ex}^{n+1} - c^{n+1,**} = 2A \cdot \left(\frac{\Delta t}{2}\right)^{p+1} + \mathcal{O}\left(\Delta t^{p+2}\right) \tag{2.105}$$

En négligeant les termes d'ordre supérieurs à $p+1$, on a alors les relations suivantes :

$$c_{Ex}^{n+1} - c^{n+1,*} = A \cdot \Delta t^{p+1} \tag{2.106}$$

$$c_{Ex}^{n+1} - c^{n+1,**} = \frac{A}{2^p} \cdot \Delta t^{p+1} \tag{2.107}$$

Où $c^{n+1,*}$ est la solution de l'équation (2.101) obtenue en un seul pas de temps et $c^{n+1,**}$ est la solution obtenue en 2 demi pas de temps. En soustrayant (2.107) de (2.106) et en réorganisant les termes, on obtient une estimation du terme A,

$$A = \frac{2^p}{\Delta t^{p+1}} \frac{c^{n+1,**} - c^{n+1,*}}{2^p - 1} \tag{2.108}$$

qui peut être réinjecté dans l'équation (2.107) :

$$c_{Ex}^{n+1} - c^{n+1,**} = \frac{c^{n+1,**} - c^{n+1,*}}{2^p - 1} + \circ\left(\Delta t^{p+2}\right) \tag{2.109}$$

Il est ainsi possible de calculer une solution extrapolée c^{n+1} de l'équation (2.101), obtenue à l'ordre $p+1$ en temps :

$$c^{n+1} = c^{n+1,**} + \frac{c^{n+1,**} - c^{n+1,*}}{2^p - 1} \tag{2.110}$$

2.4.2.1.2. Pas de temps adaptatif

L'erreur obtenue pas la solution extrapolée (2.110) est donnée par :

$$Err\left(\Delta t\right) = \left| c_{Ex}^{n+1} - c^{n+1} \right| \tag{2.111}$$

Comme c^{n+1} est une approximation d'ordre $p+2$, il en découle l'inégalité suivante :

$$Err\left(\Delta t\right) \le \left| c_{Ex}^{n+1} - c^{n+1,**} \right| \tag{2.112}$$

En combinant l'équation (2.109) et l'inégalité (2.112) et en négligeant les termes d'ordre $p+1$, on obtient un majorant de l'erreur :

$$Err\left(\Delta t\right) \le \left| \frac{c^{n+1,**} - c^{n+1,*}}{2^p - 1} \right| \tag{2.113}$$

Si l'on définit ε comme le critère de précision que l'on souhaite respecter, l'inégalité (2.113) signifie que le pas de temps Δt doit être choisi de telle façon que :

$$\left| \frac{c^{n+1,**} - c^{n+1,*}}{2^p - 1} \right| \le \varepsilon \tag{2.114}$$

Supposons que le calcul a été mené avec un pas de temps Δt_{init}. Si l'inégalité (2.114) est respectée, le calcul est alors validée et la simulation peut se poursuivre, avec un nouveau pas de temps, éventuellement plus grand que Δt_{init}. Au contraire, si cette inégalité n'est pas respectée, le calcul doit être rejeté et recommencé avec une pas de temps plus petit que Δt_{init}. Il faut donc déterminer un nouveau pas de temps Δt_{new}. Si l'on suppose que A demeure constant, l'équation (2.108) conduit au système suivant :

$$\begin{cases} A = \dfrac{2^p}{\Delta t_{init}{}^{p+1}} \dfrac{c^{n+1,**}\left(\Delta t_{init}\right) - c^{n+1,*}\left(\Delta t_{init}\right)}{2^p - 1} \\ A = \dfrac{2^p}{\Delta t_{new}{}^{p+1}} \dfrac{c^{n+1,**}\left(\Delta t_{new}\right) - c^{n+1,*}\left(\Delta t_{new}\right)}{2^p - 1} \end{cases} \tag{2.115}$$

Or pour respecter le critère de précision choisi, il faut avoir :

$$\left| \frac{c^{n+1,**}\left(\Delta t_{new}\right) - c^{n+1,*}\left(\Delta t_{new}\right)}{2^p - 1} \right| = \varepsilon \tag{2.116}$$

Ce qui permet de définir le nouveau pas de temps :

$$\Delta t_{new} = \sqrt[p+1]{\frac{\varepsilon}{\dfrac{c^{n+1,**}\left(\Delta t_{init}\right) - c^{n+1,*}\left(\Delta t_{init}\right)}{2^p - 1}}} \cdot \Delta t_{init} \tag{2.117}$$

2.4.2.1.3. Application à la cinétique chimique

L'exemple suivant représente la croissance d'une population bactérienne $[B]$ grâce à la dégradation d'un substrat carboné $[C]$ en situation aérobie avec une concentration $[O_2]$ en oxygène dissous. La croissance bactérienne est exprimée par une loi de Monod (2.118) ; la consommation de substrat par la relation (2.119) ; et l'évolution de la concentration en oxygène dissous est contrôlée par la consommation due à la biomasse vivante et par les échanges avec l'atmosphère (2.120).

$$\frac{d[B]}{dt} = \left(k_g \frac{[C]\cdot[O_2]}{\left(K_C + [C]\right)\left(K_O + [O_2]\right)} - k_r \frac{[O_2]}{\left(K_O + [O_2]\right)} \right) \cdot [B] \tag{2.118}$$

$$\frac{d[C]}{dt} = \frac{k_g}{Y} \frac{C\cdot[O_2]}{\left(K_C + [C]\right)\left(K_O + [O_2]\right)} \cdot [B] \tag{2.119}$$

$$\frac{d[O_2]}{dt} = -k_g \frac{1-Y}{Y} \frac{[C]\cdot[O_2]}{\left(K_C + [C]\right)\left(K_O + [O_2]\right)} \cdot [B] - k_r \frac{1-Y}{Y} \frac{[O_2]}{\left(K_O + [O_2]\right)} \cdot [B] + k_{exch}\left([O_2] - \left[O_2^{Eq}\right]\right) \tag{2.120}$$

Les valeurs des différents paramètres sont données dans leTableau 2-2.

Afin de tester l'efficacité de l'extrapolation de Richardson, le système différentiel composé des équations (2.118), (2.119) et (2.120) est résolu à l'aide de différentes méthodes numériques : une résolution explicite en temps, une résolution implicite en temps, un schéma implicite-explicite (Cranck-Nicholson), une méthode Runge-Kutta d'ordre 2 et une résolution par Runge-Kutta d'ordre 4. La solution de référence utilisée pour calculer les erreurs est obtenue par une méthode Runge-Kutta d'ordre 4 avec un pas de temps constant extrêmement petit (10 fois plus petit que le plus petit pas de temps utilisé par toutes les méthodes testées).

Tableau 2-2 : Valeurs des paramètres pour le test de simulation de croissance bactérienne

Parameter	Value	Signification
K_C	$1.98\cdot10^{-2}$ g.l^{-1} (a)	half life Monod coefficient for the carbonated substrate
K_O	$4.00\cdot10^{-3}$ g.l^{-1} (b)	half life Monod coefficient for the dissolved oxygen
k_g	$2.96\cdot10^{-5}$ s^{-1} (a)	consummation rate of oxygen and substrate for biomass growth
k_r	$3.47\cdot10^{-6}$ s^{-1} (a)	respiration rate of the biomass
Y	0.67 (a)	efficiency coefficient for conversion of substrate to biomass
k_{exch}	$1.16\cdot10^{-5}$ s^{-1} (c)	rate of oxygen exchange between atmosphere and liquid phase
$\left[O_2^{Eq}\right]$	$12.00\cdot10^{-3}$ g.l^{-1} (c)	dissolved oxygen concentration at equilibrium
$[B](t=0)$	10^{-3} g.l^{-1}	
$[C](t=0)$	$0.10\cdot$ g.l^{-1}	
$[O_2](t=0)$	10^{-2} g.l^{-1}	

Sources: (a) Insel et al. (2002); (b) Viotti et al. (2002); (c) Sigg et al. (2000)

La Figure 2-5 présente les résultats de ce test. Les différents temps de calculs sont obtenus, soit en changeant le pas de temps pour les méthodes à pas de temps fixe (*fixed*) soit en modifiant la valeur du critère de précision ε pour les méthodes basées sur l'extrapolation de Richardson (*Richardson*). On constate ainsi que l'extrapolation conduit bien à une solution d'ordre $p+1$ en temps : par exemple, les courbes *Runge-Kutta 2 fixed* (ordre 2 en temps) et *Explicit Richardson* sont superposées. On constate également que, à précision équivalente, il est plus rapide de faire les 3 calculs de la solution nécessaire à l'extrapolation de Richardson plutôt que de faire un seul calcul avec un pas de temps plus court.

Afin de ne pas rejeter inutilement des calculs déjà effectués, une tolérance a été rajoutée dans le critère de précision (2.114). Tant que l'erreur est inférieure à 5ε, le calcul est accepté, mais le nouveau pas de temps sera réduit pour viser une erreur égale à ε. La Figure 2-6 présente l'évolution de l'erreur estimée (2.113) et de l'erreur exacte calculée à partir de la solution de référence au cours de la simulation, pour une résolution par un schéma explicite et pour une méthode Runge-Kutta d'ordre 4, et pour 2 critères de précision relative : $\varepsilon=10^{-3}$ et $\varepsilon=10^{-5}$. On constate que l'estimateur d'erreur proposé est bien un majorant de l'erreur et que, durant toute la simulation, l'erreur exacte est bien inférieure ou égale au critère de précision choisi. Seuls 3 points, calculés par méthode Runge-Kutta 4 sont légèrement supérieurs à 10^{-3} (inférieurs à 3.10^{-3}), mais que ces points auraient été rejetés sans la tolérance introduite dans le critère de précision.

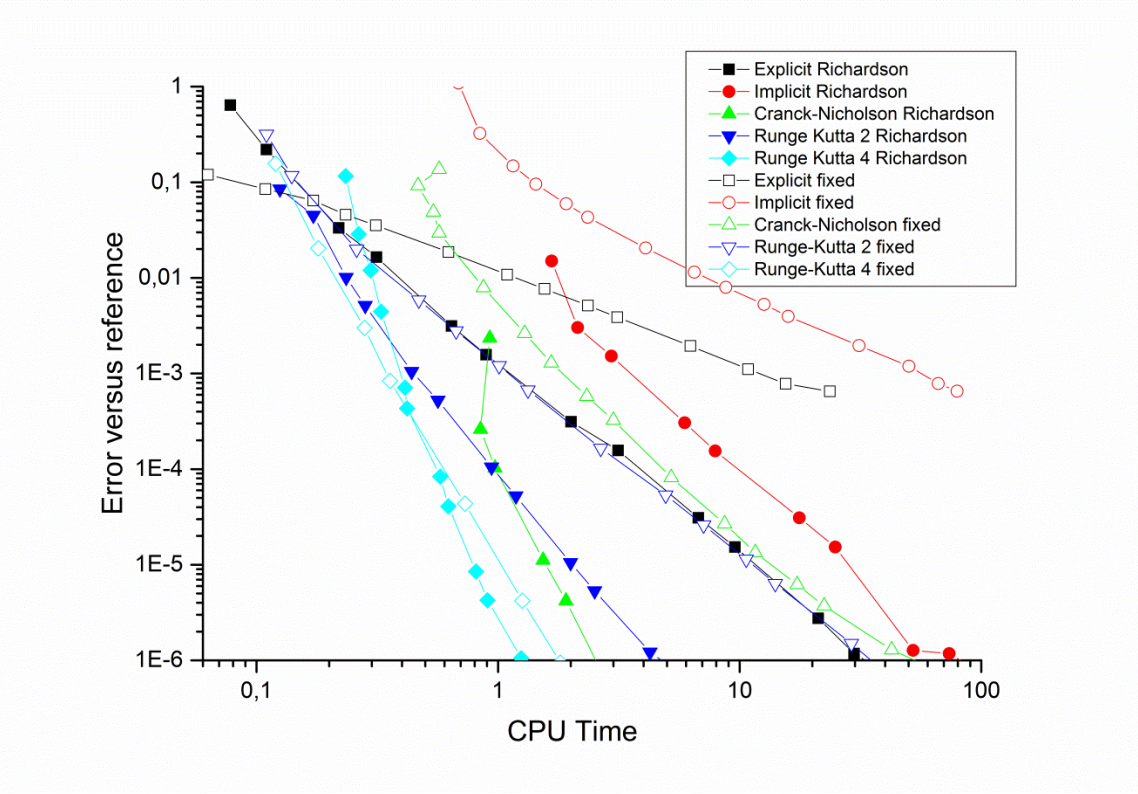


Figure 2-5 : Evolution du temps de l'erreur relative en fonction du temps de calcul nécessaire.

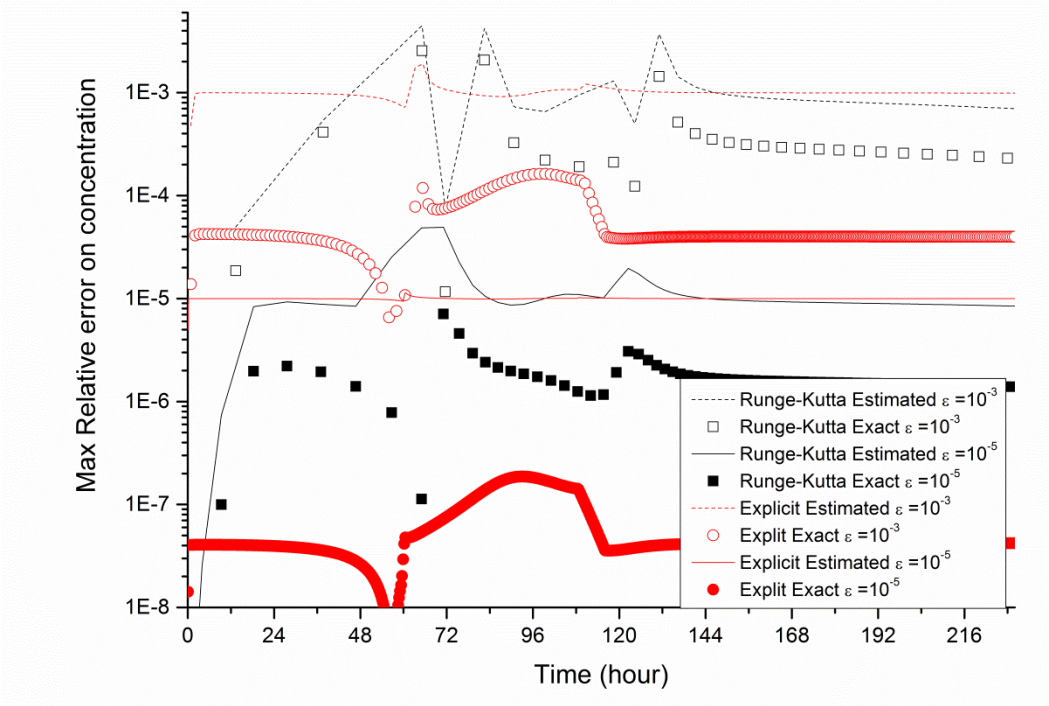


Figure 2-6 : Evolution de l'erreur estimée et de l'erreur exacte au cours de la simulation.

CONCLUSION

Le développement d'un code de transport réactif (ne devrait-on pas plutot déjà parler de suite logicielle ?) est un travail de longue haleine, dépassant les capacité d'une personne seule, comme le soulignent Steefel *et al.* [295]. Le travail présenté ici s'appuie largement sur les travaux réalisés à l'Université de Strasbourg par l'équipe *milieu poreux* (sous diverses appellations successives), au sein de l'IMFS puis du LHyGeS, pour la construction de modèles et de codes d'écoulement et de transport en milieu poreux. J'ai eu la chance et le plaisir d'encadrer ou de co-encadrer de nombreux stagiaires et doctorants qui ont apporté de riches contributions. Cependant, je pense que nous ne disposons pas au sein du LHyGES, des forces humaines suffisantes pour developer un code à vocation commerciale ou à forte diffusion (PHREEQC, HYTEC, CRUNCHFLOW…). En effet, dans le cadre d'un développement pour diffusion, la quantité de travail dédiée au contrôle, à la vérification systématique et au maintient de capacités du code est extrêmement important.

Il semble plus pertinent, compte tenu des capacités humaines actuelles de l'équipe, de maintenir une politique de diffusion restreinte, en développant des outils adaptés à chaque cas spécifique. La difficulté sera de maintenir le code suffisemment *ouvert* pour pouvoir basculer, le moment venu, vers une structure de diffusion plus large. Ce point était relativement facile à assurer tant que nous n'avions pas intégré de lois cinétiques pré-établies ni d'interface utilisateur ; mais avec le développement de ces deux points, il devient difficile de maintenir cette flexibilité dans le code.

Phénomènes

Nous avons présenté une synthèse partielle et partiale des différents phénomènes impliqués dans le transport de solutés réactifs en milieu poreux saturé ainsi que des modèles mathématiques développés pour les décrire. Cette synthèse est partielle car, comme aura pu le constater le lecteur, les phénomènes impliqués sont extrêmement nombreux et complexes [82]. Elle est de plus partiale, car, en sus des oublis, certains points ont été écarté (transport colloidal [22]) ou abordés très superficiellement (phénomènes biologiques). Ces choix ont été faits pour restreindre la présentation aux éléments déjà présents dans le modèle ou en projet d'insertion à cours terme.

Il existe de nombreux articles de synthèse récents portant sur les enjeux de l'approche multiphysique [296], sur le transfert multiphasique [52], sur le transport réactif en milieu fracturé [48], sur les problèmes d'intrusion salée et de contraste de densité [156, 157, 297], sur le transfert colloidal [22], ou sur le transport réactif en général [2, 3, 8, 77, 82, 85, 295]. On y retrouve toujours ce besoin d'intégrer l'aspect trans-disciplinaire et multiphysique spécifique à ce champ d'étude que constitue le transport réactif en milieu poreux. Cette tendance a été impulsée au début des années 2000 [8, 9, 78, 85, 181, 298] suite aux spectaculaires progrès réalisés par les outils informatiques à cette époque.

Ce travail d'intégration de phénomènes multiples au sein des différents codes de transport réactif est actuellement très avancé comme celà peut se constater à la lecture de l'article de synthèse de Steefel *et al.* [295]. Cependant, cette tâche d'intérgation est loin d'être achevée et, pour notre modèle il a été mis en évidence certains éléments à développer en priorité (voir Tableau 1) :

(i) Poursuivre la phase d'intégration des modèles de variation de porosité.

(ii) Renforcer l'intégration de modèles cinétiques, comme des modèles de précipitation-dissolution ou des modèles de biodégradation.
(iii) Développer la prise en compte des contrastes de densité.
(iv) Introduire la variation viscosité de la phase fluide en fonction de la concentration des espèces dissoutes.
(v) Prendre en compte les phénomènes d'électromigration pour pouvoir utiliser des coefficients de diffusion spécifiques à chaque espèce tout en assurant le respect de l'électroneutralité.

L'objectif de ces développements est de renforcer les liens avec les autres équipes du LHyGeS. De nombreux travaux portent sur les problèmes de diagénèse et/ou d'altération tandis que d'autres portent sur le devenir de composés d'origine anthropique (molécules phytosanitaires ou pharmaceutiques) dans les hydrosystèmes. Les phénomènes d'altération mettent en jeu des écoulements très fortement diffusifs pour lesquels la variation de porosité ainsi que la prise en compte de coefficients de diffusion spécifique à chaque espèce sont indispensables. Le transfert et la dégradation des molécules phytosanitaires et pharmaceutiques dans les hydrosystèmes demandent à être décrits à l'aide de modèles de biodégradation complexes faisant intervenir les carractéristiques physico-chimiques de l'eau (pH, oxygène…) ainsi qu'une description de la matière organique présente.

Numérique

Le code numérique développé s'appuie les travaux réalisés à l'Institut de Mécanique des Fluides et des Solides de Strasbourg et au LHyGeS [13, 108, 187, 189, 299, 300]. Les modules d'écoulement et de transport sont intégralement issus de ces travaux. Dans l'optique de développements futurs, nous pouvons nous appuier sur les travaux existants en écoulements densitaires [155, 301, 302] ou sur ceux en zone non-saturée [93, 303-305]. Les travaux numériques réalisés ont permis d'assurer un bon contrôle de l'erreur de séparation d'opérateurs et une résolution plus robuste et rapide des systèmes chimiques complexes.

En l'état actuel, le code propose une suite numérique complète intégrant les modules d'écoulement, de transport, de chimie et d'estimation de paramètres chimiques (voir Tableau 1). Le domaine physique décrit peut être un réacteur fermé, mono-, bi-, ou tri-dimensionnel ; homogène ou hétérogène physiquement et/ou chimiquement. Les phénomènes chimiques sont décrits à l'équilibre instantané et/ou sous forme cinétique, avec prise en compte de correction d'activité et du potentiel de surface.

De nombreux travaux sont en cours, avec pour finalité de proposer un code plus simple à l'usage et plus rapide (voir tableau 1). Une interface utilisateur est en développement, ainsi que l'accès automatique à des bases de données thermodynamiques. Celà permettra, à terme, une prise en main plus rapide et intuitive de la partie chimique. La parallélisation du code est en cours de développement. La structure modulaire du code, soutenue par la résolution par séparation d'opérateurs, semble extrèmement performante. Les premières tentatives de parallélisation ont été réalisées sous Open MP. Les processus de pré-processing, de post-processing et de calcul d'écoulement sont gérés sur un seul processeur. Le processus de transport est distribué sur l'ensemble des processeurs, car chaque composant est transporté indépendamment des autres. Compte tenu du nombre restreint de composants, il est fréquent de n'avoir qu'un composant transporté pour chaque processeur. La stratégie de distribution est, pour la partie transport, de faible importance. Le processus de chimie est distribué sur l'ensemble des processeurs, car la chimie de chaque maille est résolue indépendamment des autres. Dans ce cas, la stratégie de distribution joue un grand rôle dans la performance générale de la parallélisation, car le calcul de nombreuses mailles est extrèmement

rapide alors que quelques unes requièrent un temps très long. Ces premiers résultats sont prometteurs, mais nécessiteront encore de longs développements.

Comme celà est souligné dans plusieurs travaux de synthèse [76, 82, 295, 296], il me semble de plus en plus nécessaire de mettre en place des stratégies d'adaptation du pas de temps et du maillage. Le choix de séparation d'opérateurs impose, dans certains cas, des pas de temps ou d'espace assez fins pour minimiser les erreurs de séparation. Dans d'autres cas, ce sont les conditions physico-chimiques elles-même qui imposent ces efforts de calcul. La mise en place de méthodes d'adaptation en temps et en espace devrait permettre de concentrer davantage le temps de calcul sur les éléments *sensibles* de la simulation (domaines d'espace et/ou de temps ou la solution subit de fortes variations).

Applications

Les codes de transport réactif actuels sont arrivés à un niveau de maturité suffisant pour offrir des éléments de description, de compréhension et de prédiction pour des hydrosystèmes naturels complexes [2, 3, 77, 85]. Les choix de développement effectués orientent le champ d'utilisation de notre code vers les systèmes allant de l'échelle centimétrique à l'échelle kilométrique, en excluant l'échelle du pore. A ces échelles, les paramètres utilisés par les modèles sont toujours difficilement déterminables. Aux grandes échelles, pour les problèmes en milieu naturel, les hétérogénéites intrinsèques du milieu et les fluctuations des conditions imposent un calage du modèle. Comme cela a été montré par Brusseau [81], les modèles actuels sont devenu si complexes qu'il sera possible de trouver un jeu de paramètres permettant de décrire une expérience, même si les phénomènes modélisés ne sont pas adaptés.

La démarche qui me semble intéressante à développer est une approche méthodologique axée non seulement sur l'estimation des paramètres, mais aussi sur la détermination des phénomènes à inclure dans le modèle. Il s'agira de construire une démarche alliant travail expérimental et modélisation. L'objectif est d'établir une succession de conditions expérimentales conduisant à l'élaboration d'un modèle phénoménologique non équivoque. Certaines étapes de cette succession sont déjà très bien établies, mais plutôt indépendantes les unes des autres :

(i) Détermination des paramètres physiques du milieu par des mesures indépendantes : porosité par des mesures au porosimètre, perméabilité et dispersivité par des courbes d'élution de traceurs. Ce qui est moins systématique, c'est la réalisation de ces courbes d'élution à des débits différents et/ou pour des longueurs de colonne différentes pour contrôler la pertinence du traceur choisi ou vérifier l'absence de double porosité.

(ii) Construction d'isothermes de sorption pour déterminer les constantes d'équilibre et les capacités de sorption des solides (CEC ou densité de site de surface). Cependant, il est moins fréquent de réaliser ces isothermes à différentes conditions expérimentales (pH, force ionique…). C'est pourtant par ce biais que l'on pourra valider ou invalider le modèle réactionnel choisi.

(iii) Mesures de cinétiques réactionnelles, pour déterminer les constantes de vitesses réactionnelles. Encore une fois, il est assez peu fréquent de réaliser ces mesures à différentes conditions expérimentales.

(iv) Réalisation d'expériences de transport réactif en colonne de laboratoire. Ces expériences permettent l'obtention de courbes d'élution qui servent à valider un modèle complet de transport réactif. Pour ces expériences aussi, il est important de réaliser des expériences avec différentes conditions expérimentales (longueur de colonne, débit, pH, force ionique, concentration injectée…).

On constate rapidement à l'examen de cette liste que le nombre d'expériences envisagées pour l'étude d'une problématique deviendrait extrèmement important. Ceci aurait bien sûr pour conséquence un coût financier et une durée de réalisation souvent déraisonnable. L'objectif de la méthodologie envisagée sera de proposer un plan d'expériences qui soient à la fois minimales (contraintes économique et humaines) et discriminantes (valider ou invalider clairement tel ou tel choix de phénomène).

Tableau 1 : Bilan des modules dévveloppés et en cours dans le code SPECTR.
Les éléments en oranges sont en projet de développement (source du tableau Steefel *et al.* 2015).

Physique	
Dimension	
0D réacteur fermé	OUI
1D	OUI
2D	OUI
3D	OUI
Ecoulement	
Saturé	OUI
Equation richards	NON
Multiphase multiconstituant	NON
Contraste de densité	NON
Non isotherme	NON
Transport	
Advection	OUI
Diffusion	OUI
Migration électrochimique	NON
Tenseur de dispersion	Plein
Advection phase gaz	NON
Diffusion phase gaz	NON
Colloïdes	NON
Multiples continuum	NON

Géochimie	
Activité	
modèle Debye-Hückel	OUI
modèle Davies	OUI
modèle Pitzer	NON
Non isotherme	NON
Equilibre	
Complexation de surface	CCM, DLM, TLM
Echange d'ions	OUI
Precipitation-dissolution	OUI
Solution solide	OUI
Précipitation de surface	OUI
Fractionnement isotopique	OUI
Echange eau-gaz	OUI
Cinétique	
Precipitation-dissolution	OUI
Fractionnement isotopique	OUI
Cinétique en phase aqueuse	OUI
Nucléation minérale	NON
Décroissance radioactive	OUI
Microorganismes	
Cinétique Monod	OUI
Thermodynamique	NON
Croissance biomasse	NON

Phénomènes couplés	
Variation de porosité-perméabilité	OUI
Couplage chaleur - réactions	NON
Couplage déformation - compactage	NON
Réaction consommation de phase (H2O)	NON
Transport dans la double couche électrostatique	NON
Schéma numérique	
Séparation d'opérateurs	SNIA
Approche globale	NON
Transport à grand nombre de Péclet	OUI
Discrétisation spatiale	EF discontinus / mixtes hybrides
Discrétisation temporelle	Explicite / Implicite
Calcul	
Parallélisation	en cours
Estimation de paramètres	OUI
Interface utilisateur	en cours
Open source	OUI

BIBLIOGRAPHIE

1. de Marsily, G., *An overview of the world's water resources problems in 2050.* Ecohydrology & Hydrobiology, 2007. **7**(2): p. 147-155.

2. Yaron, B., I. Dror, and B. Berkowitz, *Contaminant geochemistry-a new perspective.* Naturwissenschaften, 2010. **97**(1): p. 1-17.

3. Saripalli, K.P., et al., *Changes in hydrologic properties of aquifer media due to chemical reactions: A review.* Critical Reviews in Environmental Science and Technology, 2001. **31**(4): p. 311-349.

4. Li, L., C.H. Benson, and E.M. Lawson, *Modeling porosity reductions caused by mineral fouling in continuous-wall permeable reactive barriers.* Journal of Contaminant Hydrology, 2006. **83**(1–2): p. 89-121.

5. Lichtner, P.C. and N. Waber, *Redox front geochemistry and weathering: theory with application to the Osamu Utsumi uranium mine, Po¤os de Caldas, Brazil.* Journal of Geochemical Exploration, 1992. **45**(1-3): p. 521-564.

6. Loyaux-Lawniczak, S., P. Lecomte, and J.J. Ehrhardt, *Behavior of hexavalent chromium in a polluted groundwater: Redox processes and immobilization in soils.* Environmental Science & Technology, 2001. **35**(7): p. 1350-1357.

7. Wanner, C., et al., *A chromate-contaminated site in southern Switzerland – Part 1: Site characterization and the use of Cr isotopes to delineate fate and transport.* Applied Geochemistry, 2012. **27**(3): p. 644-654.

8. Steefel, C.I., *New directions in hydrogeochemical transport modeling: Incorporating multiple kinetic and equilibrium reaction pathways.* Computational methods in water resources, 2000. **1**: p. 331-338.

9. Parlange, M.B., et al., *Editorial: Future of Water Resources Research.* Water Resources Research, 2005. **41**(1): p. 1-2.

10. Valocchi, A.J., R.L. Street, and P.V. Roberts, *Transport of ion-exchanging solutes in groundwater: Chromatographic theory and field simulation.* Water Resour.Res., 1981. **17**: p. 1517-1527.

11. Walter, A.L., et al., *Modeling of multicomponent reactive transport in groundwater. 2. Metal mobility in aquifers impacted by acidic mine tailings discharge.* Water Resources Research, 1994. **30**(11): p. 3149-3158.

12. De Windt, L., S. Leclercq, and J. van der Lee. *Assessing the durability of nuclear glass with respect to silica controlling processes in a clayey underground disposal*. in *29th International Symposium on the Scientific Basis for Nuclear Waste Management XXIX*. 2005. Ghent; Belgium: Materials Research Society Symposium Proceedings.

13. Hoteit, H., P. Ackerer, and R. Mose, *Nuclear waste disposal simulations: Couplex test cases.* Computational Geosciences, 2004. **8**(2): p. 99-124.

14. Tompson, A.F.B., et al., *On the evaluation of groundwater contamination from underground nuclear tests.* Environmental Geology, 2002. **42**(2-3): p. 235-247.

15. Andre, L., et al., *Numerical modeling of fluid-rock chemical interactions at the supercritical CO2-liquid interface during CO2 injection into a carbonate reservoir, the Dogger aquifer (Paris Basin, France).* Energy Conversion and Management, 2007. **48**(6): p. 1782-1797.

16. Kang, Q., et al., *Pore scale modeling of reactive transport involved in geologic CO2 sequestration.* Transport in Porous Media, 2010. **82**(1): p. 197-213.

17. Navarre-Sitchler, A.K., et al., *Elucidating geochemical response of shallow heterogeneous aquifers to CO2 leakage using high-performance computing: Implications for monitoring of CO2 sequestration.* Advances in Water Resources, 2013. **53**(0): p. 45-55.

18. Pruess, K., et al., *Code intercomparison builds confidence in numerical simulation models for geologic disposal of CO2.* Energy, 2004. **29**(9-10): p. 1431-1444.

19. Regnault, O., et al., *Etude experimentale de la reactivite du CO2 supercritique vis-a-vis de phases minerales pures. Implications pour la sequestration geologique de CO2.* Comptes Rendus Geosciences, 2005. **337**(15): p. 1331-1339.

20. Bai, C. and Y. Li, *Modeling the transport and retention of nC60 nanoparticles in the subsurface under different release scenarios.* Journal of Contaminant Hydrology, 2012. **136–137**: p. 43-55.

21. He, F., et al., *Transport of carboxymethyl cellulose stabilized iron nanoparticles in porous media: Column experiments and modeling.* Journal of Colloid and Interface Science, 2009. **334**(1): p. 96-102.

22. Kanti Sen, T. and K.C. Khilar, *Review on subsurface colloids and colloid-associated contaminant transport in saturated porous media.* Advances in Colloid and Interface Science, 2006. **119**(2-3): p. 71-96.

23. Kurwadkar, S., et al., *Evaluation of leaching potential of three systemic neonicotinoid insecticides in vineyard soil.* Journal of Contaminant Hydrology, 2014. **170**: p. 86-94.

24. Rodríguez-Liébana, J.A., M.D. Mingorance, and A. Peña, *Thiacloprid adsorption and leaching in soil: Effect of the composition of irrigation solutions.* Science of The Total Environment, 2018. **610–611**: p. 367-376.

25. Whiting, S.A., et al., *A multi-year field study to evaluate the environmental fate and agronomic effects of insecticide mixtures.* Science of The Total Environment, 2014. **497–498**: p. 534-542.

26. Luque-Espinar, J.A., et al., *Seasonal occurrence and distribution of a group of ECs in the water resources of Granada city metropolitan areas (South of Spain): Pollution of raw drinking water.* Journal of Hydrology, 2015. **531, Part 3**: p. 612-625.

27. Peake, B.M., et al., *5 - Impact of pharmaceuticals on the environment*, in *The Life-Cycle of Pharmaceuticals in the Environment*. 2016, Woodhead Publishing. p. 109-152.

28. Yabusaki, S., et al., *Multicomponent reactive transport in an in situ zero-valent iron cell.* Environmental Science and Technology, 2001. **35**(7): p. 1493-1503.

29. Jacques, D., et al., *Operator-splitting errors in coupled reactive transport codes for transient variably saturated flow and contaminant transport in layered soil profiles.* Journal of Contaminant Hydrology, 2006. **88**(3-4): p. 197-218.

30. Gurban, I., et al., *Uranium transport around the reactor zone at Bangombé and Okélobondo (Oklo): Examples of hydrogeological and geochemical model integration and data evaluation.* Journal of Contaminant Hydrology, 2003. **61**(1-4): p. 247-264.

31. De Windt, L., R. Badreddine, and V. Lagneau, *Long-term reactive transport modelling of stabilized/solidified waste: from dynamic leaching tests to disposal scenarios.* Journal of Hazardous Materials, 2007. **139**(3): p. 529-536.

32. Guimaraes, L.D.N., A. Gens, and S. Olivella, *Coupled thermo-hydro-mechanical and chemical analysis of expansive clay subjected to heating and hydration.* Transport in Porous Media, 2007. **66**(3): p. 341-372.

33. Lemaire, T., C. Moyne, and D. Stemmelen, *Modelling of electro-osmosis in clayey materials including pH effects.* Physics and Chemistry of the Earth, Parts A/B/C, 2007. **32**(1-7): p. 441-452.

34. Ojeda, E., et al., *Evaluation of relative importance of different microbial reactions on organic matter removal in horizontal subsurface-flow constructed wetlands using a 2D simulation model.* Ecological Engineering, 2008. **34**(1): p. 65-75.

35. Lu, C., et al., *Effects of density and mutual solubility of a –brine system on storage in geological formations: "Warm" vs. "cold" formations.* Advances in Water Resources, 2009. **32**(12): p. 1685-1702.

36. Mangeret, A., L. De Windt, and P. Crançon, *Reactive transport modelling of groundwater chemistry in a chalk aquifer at the watershed scale.* Journal of Contaminant Hydrology, 2012. **138–139**(0): p. 60-74.

37. Cheng, H.P. and G.T. Yeh, *Development and demonstrative application of a 3-D numerical model of subsurface flow, heat transfer, and reactive chemical transport: 3DHYDROGEOCHEM.* Journal of Contaminant Hydrology, 1998. **34**(1-2): p. 47-83.

38. Charbeneau, R.J., *Groundwater contaminant transport with adsorption and ion exchange chemistry: method of characteristics for the case without dispersion.* Water Resources Research, 1981. **17**(3): p. 705-713.

39. Jennings, A.A., D.J. Kirkner, and T.L. Theis, *Multicomponent equilibrium chemistry in groundwater quality models.* Water Resour.Res., 1982. **18**: p. 1089-1096.

40. Steefel, C.I. and A.C. Lasaga, *A coupled model for transport of multiple chemical species and kinetic precipitation/dissolution reactions with application to reactive flow in single phase hydrothermal systems.* American Journal of Science, 1994. **294**(5): p. 529-592.

41. Steefel, C.I. and P.C. Lichtner, *Multicomponent reactive transport in discrete fractures: I. Controls on reaction front geometry.* Journal of Hydrology, 1998. **209**(1-4): p. 186-199.

42. Steefel, C.I. and P.C. Lichtner, *Multicomponent reactive transport in discrete fractures: II: Infiltration of hyperalkaline groundwater at Maqarin, Jordan, a natural analogue site.* Journal of Hydrology, 1998. **209**(1-4): p. 200-224.

43. Yabusaki, S.B., C.I. Steefel, and B.D. Wood, *Multidimensional, multicomponent, subsurface reactive transport in nonuniform velocity fields: code verification using an advective reactive streamtube approach.* Journal of Contaminant Hydrology, 1998. **30**(3-4): p. 299-331.

44. De Windt, L., D. Pellegrini, and J. van der Lee, *Reactive transport modelling of a spent fuel repository in a stiff clay formation considering excavation damaged zones.* Radiochimica Acta, 2004. **92**(9-11): p. 841-848.

45. Trotignon, L., et al., *Design of a 2-D cementation experiment in porous medium using numerical simulation.* Oil and Gas Science and Technology, 2005. **60**(2): p. 307-318.

46. van der Lee, J., E. Ledoux, and G. de Marsily, *Modeling of colloidal uranium transport in a fractured medium.* Journal of Hydrology, 1992. **139**(1-4): p. 135-158.

47. Steefel, C.I. and P.C. Lichtner, *Diffusion and reaction in rock matrix bordering a hyperalkaline fluid-filled fracture.* Geochimica et Cosmochimica Acta, 1994. **58**(17): p. 3595-3612.

48. MacQuarrie, K.T.B. and K.U. Mayer, *Reactive transport modeling in fractured rock: A state-of-the-science review.* Earth-Science Reviews, 2005. **72**(3-4): p. 189-227.

49. Lichtner, P.C., *Continuum formulation of multicomponent-multiphase reactive transport.* Reviews in Mineralogy, 1996. **34**.

50. Saaltink, M.W., et al., *Simulating reactive transport in time dependent multiphase flow problems.* Radiochimica Acta, 2004. **92**(9-11): p. 835-839.

51. Saaltink, M.W., et al., *RETRASO, a code for modeling reactive transport in saturated and unsaturated porous media.* Geologica Acta, 2004. **2**(3): p. 235-251.

52. Helmig, R., et al., *Model coupling for multiphase flow in porous media.* Advances in Water Resources, 2013. **51**: p. 52-66.

53. Carnahan, C.L. and J.S. Remer, *Nonequilibrium and equilibrium sorption with a linear sorption isotherm during mass transport through an infinite porous medium: Some analytical solutions.* Journal of Hydrology, 1984. **73**(3-4): p. 227-258.

54. Spurlock, F.C., K. Huang, and M.T. van Genuchten, *Isotherm nonlinearity and nonequilibrium sorption effects on transport of fenuron and monuron in soil columns.* Environmental Science and Technology, 1995. **29**(4): p. 1000-1007.

55. Appelo, C.A.J., *Some calculations on multicomponent transport with cation exchange in aquifers.* Ground Water, 1994. **32**(6): p. 968-975.

56. Steefel, C.I., et al., *Cesium migration in Hanford sediment: A multisite cation exchange model based on laboratory transport experiments.* Journal of Contaminant Hydrology, 2003. **67**(1-4): p. 219-246.

57. Walsh, M.P., et al., *Precipitation and dissolution of solids attending flow through porous media.* AIChE Journal, 1984. **30**: p. 317-328.

58. Lefevre, F., M. Sardin, and D. Schweich, *Migration of strontium in clayey and calcareous sandy soil: Precipitation and ion exchange.* Journal of Contaminant Hydrology, 1993. **13**(1-4): p. 215-229.

59. De Windt, L., et al., *Intercomparison of reactive transport models applied to UO2 oxidative dissolution and uranium migration.* Journal of Contaminant Hydrology, 2003. **61**(1-4): p. 303-312.

60. Matsunaga, T., et al., *Redox chemistry of iron and manganese minerals in river-recharged aquifers: A model interpretation of a column experiment.* Geochimica et Cosmochimica Acta, 1993. **57**(8): p. 1691-1704.

61. van der Lee, J., *CHESS another speciation and surface complexation computer code.*, E.d.M.d. Paris, Editor. 1993: Fontainebleau. p. 85.

62. Aggarwal, M. and J. Carrayrou, *Parameter estimation for reactive transport by a Monte-Carlo approach.* AIChE Journal, 2006. **52**(6): p. 2281-2289.

63. Aggarwal, M., M. Cheikh Anta Ndiaye, and J. Carrayrou, *Parameters estimation for reactive transport: A way to test the validity of a reactive model.* Physics and Chemistry of the Earth, Parts A/B/C, 2007. **32**(1-7): p. 518-529.

64. Lichtner, P.C. and J.W. Carey, *Incorporating solid solutions in reactive transport equations using a kinetic discrete-composition approach.* Geochimica et Cosmochimica Acta, 2006. **70**(6): p. 1356-1378.

65. Shao, H., et al., *Modeling reactive transport in non-ideal aqueous–solid solution system.* Applied Geochemistry, 2009. **24**(7): p. 1287-1300.

66. Wanko, A., et al., *Simulation of biodegradation in infiltration seepage - Model development and hydrodynamic calibration.* Water, Air, and Soil Pollution, 2006. **177**(1-4): p. 19-43.

67. Salvage, K.M. and G.-T. Yeh, *Development and application of a numerical model of kinetic and equilibrium microbiological and geochemical reactions (BIOKEMOD).* Journal of Hydrology, 1998. **209**(1–4): p. 27-52.

68. Tebes-Stevens, C., et al., *Multicomponent transport with coupled geochemical and microbiological reactions: model description and example simulations.* Journal of Hydrology, 1998. **209**(1–4): p. 8-26.

69. Nowack, B., et al., *Verification and intercomparison of reactive transport codes to describe root-uptake.* Plant and Soil, 2006. **285**(1-2): p. 305-321.

70. Roose, T., et al., *Verification and intercomparison of reactive transport codes to describe root-uptake.* Plant and Soil, 2007. **301**(1-2): p. 327.

71. Chen, J.-S. and C.-W. Liu, *Numerical simulation of the evolution of aquifer porosity and species concentrations during reactive transport.* Computers & Geosciences, 2002. **28**(4): p. 485-499.

72. Emmanuel, S. and B. Berkowitz, *Mixing-induced precipitation and porosity evolution in porous media.* Advances in Water Resources, 2005. **28**(4): p. 337-344.

73. Katz, G.E., et al., *Experimental and modeling investigation of multicomponent reactive transport in porous media.* Journal of Contaminant Hydrology, 2011. **120-121**(C): p. 27-44.

74. Le Gallo, Y., O. Bildstein, and E. Brosse, *Coupled reaction-flow modeling of diagenetic changes in reservoir permeability, porosity and mineral compositions.* Journal of Hydrology, 1998. **209**(1–4): p. 366-388.

75. Chen, J.-S., et al., *Effects of mechanical dispersion on the morphological evolution of a chemical dissolution front in a fluid-saturated porous medium.* Journal of Hydrology, 2009. **373**(1–2): p. 96-102.

76. Zhao, C., L.B. Reid, and K. Regenauer-Lieb, *Some fundamental issues in computational hydrodynamics of mineralization: A review.* Journal of Geochemical Exploration, 2012. **112**: p. 21-34.

77. Steefel, C.I., D.J. DePaolo, and P.C. Lichtner, *Reactive transport modeling: An essential tool and a new research approach for the Earth sciences.* Earth and Planetary Science Letters, 2005. **240**(3-4): p. 539-558.

78. Steefel, C. and P. Van Cappellen, *Preface: Reactive transport modeling of natural systems.* Journal of Hydrology, 1998. **209**(1–4): p. 1-7.

79. Bredehoeft, J.D. and L.F. Konikow, *Ground-Water Models: Validate or Invalidate.* Ground Water, 2012. **50**(4): p. 493-495.

80. Konikow, L.F. and J.D. Bredehoeft, *Ground-water models cannot be validated.* Advances in Water Resources, 1992. **15**(1): p. 75-83.

81. Brusseau, M.L., *Non-ideal transport of reactive solutes in heterogeneous porous media: 3. model testing and data analysis using calibration versus prediction.* Journal of Hydrology, 1998. **209**(1–4): p. 147-165.

82. Miller, C.T., et al., *Numerical simulation of water resources problems: Models, methods, and trends.* Advances in Water Resources, 2013. **51**(0): p. 405-437.

83. Yeh, G.T. and V.S. Tripathi, *A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components.* Water Resour.Res., 1989. **25**: p. 93-108.

84. Carrayrou, J., et al., *Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems-the MoMaS benchmark case.* Computational Geosciences, 2010. **14**(3): p. 483-502.

85. van der Lee, J. and L. De Windt, *Present state and future directions of modeling of geochemistry in hydrogeological systems.* Journal of Contaminant Hydrology, 2001. **47**(24): p. 265-282.

86. Carrayrou, J., *Modélisation du transport de solutés réactifs en milieu poreux saturé*. 2001, Université Louis Pasteur Strasbourg I. p. 248.

87. Carrayrou, J., R. Mosé, and P. Behra, *Modelling reactive transport in porous media: iterative scheme and combination of discontinuous and mixed-hybrid finite elements.* C.R.Acad.Sci., Ser.II Univers, 2003. **331**(3): p. 211-216.

88. Valocchi, A.J. and M. Malmstead, *Accuracy of operator splitting for advection-dispersion-reaction problems.* Water Resour.Res., 1992. **28**: p. 1471-1476.

89. Carrayrou, J., R. Mose, and P. Behra, *Operator-splitting procedures for reactive transport and comparison of mass balance errors.* Journal of Contaminant Hydrology, 2004. **68**(3): p. 239-268.

90. Fahs, M., et al., *On the efficiency of the direct substitution approach for reactive transport problems in porous media.* Water, Air, and Soil Pollution, 2008. **193**(1-4): p. 299-308.

91. Westall, J.C., J.L. Zachary, and F.M.M. Morel, *MINEQL: a computer program for the calculation of chemical equilibrium composition of aqueous system.*, R.M.P. Laboratory, Editor. 1976: Cambridge. p. 91.

92. Carrayrou, J., R. Mosé, and P. Behra, *New efficient algorithm for solving thermodynamic chemistry.* AIChE Journal, 2002. **48**(4): p. 894-904.

93. Belfort, B., J. Carrayrou, and F. Lehmann, *Implementation of Richardson extrapolation in an efficient adaptive time stepping method: applications to reactive transport and unsaturated flow in porous media.* Transport in Porous Media, 2007. **69**: p. 123-138.

94. Westall, J.C., *FITEQL ver. 2.1*. 1982: Corvallis.

95. Bueno, M., et al., *Dynamic sorptive behavior of tributyltin on quartz sand at low concentration levels: Effect of pH, flow rate, and monovalent cations.* Environ.Sci.Technol., 1998. **32**(24): p. 3919-3925.

96. Majdalani, S., et al., *Reactive transport parameter estimation: Genetic algorithm vs. monte carlo approach.* AIChE Journal, 2009. **55**(8): p. 1959-1968.

97. Carrayrou, J., et al. *GdR Momas - Benchmark reactive transport.* http://www.gdrmomas.org/Ex_qualif/Geochimie/Documents/Benchmark-MoMAS.pdf, 2008.

98. Carrayrou, J., M. Kern, and P. Knabner, *Reactive transport benchmark of MoMaS.* Computational Geosciences, 2010. **14**(3): p. 385-392.

99. Carrayrou, J., *Looking for some reference solutions for the reactive transport benchmark of MoMaS with SPECY.* Computational Geosciences, 2010. **14**(3): p. 393-403.

100. Ackerer, P., *Preface: Special issue on simulations of reactive transport: Results of the MoMaS benchmarks.* Computational Geosciences, 2010. **14**(3): p. 383.

101. Lagneau, V. and J. van der Lee, *HYTEC results of the MoMas reactive transport benchmark.* Computational Geosciences, 2010. **14**(3): p. 435-449.

102. Amir, L. and M. Kern, *A global method for coupling transport with chemistry in heterogeneous porous media.* Computational Geosciences, 2010. **14**(3): p. 465-481.

103. de Dieuleveult, C. and J. Erhel, *A global approach to reactive transport: Application to the MoMas benchmark.* Computational Geosciences, 2010. **14**(3): p. 451-464.

104. Mayer, K.U. and K.T.B. MacQuarrie, *Solution of the MoMaS reactive transport benchmark with MIN3P-model formulation and simulation results.* Computational Geosciences, 2010. **14**(3): p. 405-419.

105. Hoffmann, J., S. Kräutle, and P. Knabner, *A parallel global-implicit 2-D solver for reactive transport problems in porous media based on a reduction scheme and its application to the MoMaS benchmark problem.* Computational Geosciences, 2010. **14**(3): p. 421-433.

106. Hoteit, H., et al., *New two-dimensional slope limiters for discontinuous Galerkin methods on arbitrary meshes.* International Journal for Numerical Methods in Engineering, 2004. **61**(14): p. 2566-2593.

107. Mose, R., P. Ackerer, and G. Chavent. *Application of the mixed hybrid finite element approximation in a three - dimensional model for groundwater flow and quality modelling*.

108. Mose, R., et al., *Application of the mixed hybrid finite element approximation in a groundwater flow model: luxury or necessity?* Water Resources Research, 1994. **30**(11): p. 3001-3012.

109. Lichtner, P.C., *Continuum model for simultaneous chemical reactions and mass transport in hydrothermal systems.* Geochimica et Cosmochimica Acta, 1985. **49**(3): p. 779-800.

110. Rubin, J. and R.V. James, *Dispersion-affected transport of reacting solutes in saturated porous media: Galerkin method applied to equilibrium-controlled exchange in unidirectional steady water flow.* Water Resour.Res., 1973. **9**: p. 1332-1356.

111. Miller, W. and L.V. Benson, *Simulation of solute transport in a chemically reactive heterogeneous system: model development and application.* Water Resour.Res., 1983. **19**: p. 381-391.

112. Cederberg, A., R.L. Street, and J.O. Leckie, *A groundwater mass transport and equilibrium chemistry model for multicomponent systems.* Water Resour.Res., 1985. **21**: p. 1095-1104.

113. Lichtner, P.C., *The quasi-stationary state approximation to coupled mass transport and fluid-rock interaction in a porous medium.* Geochimica et Cosmochimica Acta, 1988. **52**(1): p. 143-165.

114. Lichtner, P.C. *Principles and practice of reactive transport modeling*. 1995.

115. Ouvrard, S., M.O. Simonnot, and M. Sardin, *Reactive behavior of natural manganese oxides toward the adsorption of phosphate and arsenate.* Industrial and Engineering Chemistry Research, 2002. **41**(11): p. 2785-2791.

116. Chilakapati, A., et al., *Groundwater flow, multicomponent transport and biogeochemistry: development and application of a coupled process model.* Journal of Contaminant Hydrology, 2000. **43**(3-4): p. 303-325.

117. Salvage, K.M. and G.T. Yeh, *Development and application of a numerical model of kinetic and equilibrium microbiological and geochemical reactions (BIOKEMOD).* Journal of Hydrology, 1998. **209**(1-4): p. 27-52.

118. Bear, J., *Dynamics of fluids in porous media.* 1972, New-York.

119. de Marsily, G., *Hydrogéologie quantitative*. 1981, Paris. 215.

120. Soler, J.M. and A.C. Lasaga, *An advection–dispersion–reaction model of bauxite formation.* Journal of Hydrology, 1998. **209**(1–4): p. 311-330.

121. Shen, L. and Z. Chen, *Critical review of the impact of tortuosity on diffusion.* Chemical Engineering Science, 2007. **62**(14): p. 3748-3755.

122. Boudreau, B.P., *The diffusive tortuosity of fine-grained unlithified sediments.* Geochimica et Cosmochimica Acta, 1996. **60**(16): p. 3139-3142.

123. Darcy, H., *Les fontaines publiques de la ville de Dijon*. 1856, Paris: Victor Dalmont. 647.

124. Zhao, C., B.E. Hobbs, and A. Ord, *Fundamentals of computational geoscience: Numerical methods and algorithms*. 2009, Springer Berlin Heidelberg. p. 1-257.

125. Homand F., D.P., *Manuel de mécanique des roches*. Vol. 1. 2000, Paris: Les Presses de l'Ecole des Mines.

126. Lichtner, P.C., E.H. Oelkers, and H.C. Helgeson, *Interdiffusion with multiple precipitation/dissolution reactions: Transient model and the steady-state limit.* Geochimica et Cosmochimica Acta, 1986. **50**(9): p. 1951-1966.

127. Merino, E. and T. Dewers, *Implications of replacement for reaction–transport modeling.* Journal of Hydrology, 1998. **209**(1–4): p. 137-146.

128. *Handbook of chemistry and physics*. 58 ed. 1978: CRC press.

129. Carrier, W.D., *Goodbye, Hazen; Hello, Kozeny-Carman.* Journal of geotechnical and geoenvironmental engineering, 2003. **129**(11).

130. Carman, P.C., *Flow of gases through porous media*. 1956, New York: Academic press.

131. Cochepin, B., et al., *Approaches to modelling coupled flow and reaction in a 2D cementation experiment.* Advances in Water Resources, 2008. **31**(12): p. 1540-1551.

132. Quispe, J.R., R.E. Rozas, and P.G. Toledo, *Permeability–porosity relationship from a geometrical model of shrinking and lattice Boltzmann and Monte Carlo simulations of flow in two-dimensional pore networks.* Chemical Engineering Journal, 2005. **111**(2–3): p. 225-236.

133. Acharya, R.C., S.E.A.T.M. van der Zee, and A. Leijnse, *Porosity–permeability properties generated with a new 2-parameter 3D hydraulic pore-network model for consolidated and unconsolidated porous media.* Advances in Water Resources, 2004. **27**(7): p. 707-723.

134. Zinszner B., P.F.M., *A Geoscientist's Guide to Petrophysics*, in *A Geoscientist's Guide to Petrophysics*, E. Technip, Editor. 2007.

135. Bildstein, O., et al., *Modelling iron–clay interactions in deep geological disposal conditions.* Physics and Chemistry of the Earth, Parts A/B/C, 2006. **31**(10–14): p. 618-625.

136. Salles, J., J.F. Thovert, and P.M. Adler, *Reconstructed porous media and their application to fluid flow and solute transport.* Journal of Contaminant Hydrology, 1993. **13**(1-4): p. 3-22.

137. Navarre-Sitchler, A., et al., *A reactive-transport model for weathering rind formation on basalt.* Geochimica et Cosmochimica Acta, 2011. **75**(23): p. 7644-7667.

138. Harleman, D.R.F., Melhorn,P.F.,Rumer,R.R., *Dispersion-permeability correlaction in porous media.* Journal of the Hydraulics Division Proceedings of the American Society of Civil Engineers, 1963. **89**(HY 2): p. 67-85.

139. Salles, J., et al., *Taylor dispersion in porous media. Determination of the dispersion tensor.* Physics of Fluids A, 1992. **5**(10): p. 2348-2376.

140. Canals, M. and J.D. Meunier, *A model for porosity reduction in quartzite reservoirs by quartz cementation.* Geochimica et Cosmochimica Acta, 1995. **59**(4): p. 699-709.

141. Kieffer, B., et al., *An experimental study of the reactive surface area of the Fontainebleau sandstone as a function of porosity, permeability, and fluid flow rate.* Geochimica et Cosmochimica Acta, 1999. **63**(21): p. 3525-3534.

142. Merino, E., P. Ortoleva, and P. Strickholm, *Generation of evenly-spaced pressure-solution seams during (late) diagenesis: A kinetic theory.* Contributions to Mineralogy and Petrology, 1983. **82**(4): p. 360-370.

143. Chen, Z.-S., et al., *Numerical investigation on the thermal non-equilibrium in low-velocity reacting flow within porous media.* International Journal of Heat and Mass Transfer, 2014. **77**: p. 585-599.

144. Jeong, S.-M., et al., *Direct numerical simulation of CO2 hydrate dissociation in pore-scale flow by applying CFD method.* International Journal of Heat and Mass Transfer, 2017. **107**: p. 300-306.

145. Liu, Z. and H. Wu, *Pore-scale study on flow and heat transfer in 3D reconstructed porous media using micro-tomography images.* Applied Thermal Engineering, 2016. **100**: p. 602-610.

146. Nagendra, K. and D.K. Tafti, *A sub-pore model for multi-scale reaction–diffusion problems in porous media.* International Journal of Heat and Mass Transfer, 2015. **84**: p. 463-474.

147. Peng, X.F. and H.L. Wu, *14 - Pore-scale transport phenomena in porous media*, in *Transport Phenomena in Porous Media III*, D.B. Ingham and I. Pop, Editors. 2005, Pergamon: Oxford. p. 366-398.

148. Pichler, C., R. Traxl, and R. Lackner, *Power-law scaling of thermal conductivity of highly porous ceramics.* Journal of the European Ceramic Society, 2015. **35**(6): p. 1933-1941.

149. Prat, M., *Recent advances in pore-scale models for drying of porous media.* Chemical Engineering Journal, 2002. **86**(1–2): p. 153-164.

150. Quintard, M., M. Kaviany, and S. Whitaker, *Two-medium treatment of heat transfer in porous media: numerical results for effective properties.* Advances in Water Resources, 1997. **20**(2–3): p. 77-94.

151. Quintard, M. and S. Whitaker, *Local thermal equilibrium for transient heat conduction: theory and comparison with numerical experiments.* International Journal of Heat and Mass Transfer, 1995. **38**(15): p. 2779-2796.

152. Teruel, F.E. and L. Díaz, *Calculation of the interfacial heat transfer coefficient in porous media employing numerical simulations.* International Journal of Heat and Mass Transfer, 2013. **60**: p. 406-412.

153. Herbert, A.W., C.P. Jackson, and D.A. Lever, *Coupled groundwater flow and solute transport with fluid density strongly dependent upon concentration.* Water Resources Research, 1988. **24**(10): p. 1781-1795.

154. Ophori, D.U., *Flow of groundwater with variable density and viscosity, Atikokan Research Area, Canada.* Hydrogeology Journal, 1998. **6**(2): p. 193-203.

155. Younes, A., et al., *Modelling variable density flow problems in heterogeneous porous media using the method of lines and advanced spatial discretization methods.* Mathematics and Computers in Simulation, 2011. **81**(10): p. 2346-2355.

156. Werner, A.D., et al., *Seawater intrusion processes, investigation and management: Recent advances and future challenges.* Advances in Water Resources, 2013. **51**(0): p. 3-26.

157. Diersch, H.J.G. and O. Kolditz, *Variable-density flow and transport in porous media: approaches and challenges.* Advances in Water Resources, 2002. **25**(8–12): p. 899-944.

158. IAPWS, *Revised Release on the IAPWS Formulation 1995 for the Thermodynamic Properties of Ordinary Water Substance for General and Scientific Use*. 2009, The International Association for the Properties of Water and Steam. p. 18.

159. Tanaka, M., et al., *Recommended table for the density of water between 0 °C and 40 °C based on recent experimental reports.* Metrologia, 2001. **38**(4): p. 301.

160. Diersch, H.J.G., *FEFLOW Reference Manual*. 2009, DHI-WASY GmbH: Berlin.

161. Viswanath, D.S., Ghosh, T., Prasad, D.H.L., Dutt, N.V.K., Rani, K.Y. , *Viscosity of Liquids, Theory, Estimation, Experiment, and Data*. 2007: Springer Netherlands.

162. Matsen, J.M. and E.F. Johnson, *Extension of the Andrade Equation for Viscosity at the Normal Melting Point.* Journal of Chemical & Engineering Data, 1960. **5**(4): p. 531-533.

163. Lencka, M.M., et al., *Modeling Viscosity of Multicomponent Electrolyte Solutions.* International Journal of Thermophysics, 1998. **19**(2): p. 367-378.

164. Onsager, L. and R.M. Fuoss, *Irreversible Processes in Electrolytes. Diffusion, Conductance and Viscous Flow in Arbitrary Mixtures of Strong Electrolytes.* The Journal of Physical Chemistry, 1931. **36**(11): p. 2689-2778.

165. Lichtner, P.C. and Q. Kang, *Upscaling pore-scale reactive transport equations using a multiscale continuum formulation.* Water Resources Research, 2007. **43**(12).

166. Chevalier, S. and O. Banton, *Modelling of heat transfer with the random walk method. Part 1. Application to thermal energy storage in porous aquifers.* Journal of Hydrology, 1999. **222**(1–4): p. 129-139.

167. Kaviany, M., *Principles of heat transfer in porous media*. 1991, United States: New York, NY (United States); Springer-Verlag New York Inc.

168. Kim, S.J. and J.M. Hyun, *5 - A porous medium approach for the thermal analysis of heat transfer devices*, in *Transport Phenomena in Porous Media III*, D.B. Ingham and I. Pop, Editors. 2005, Pergamon: Oxford. p. 120-146.

169. Niessner, J. and S.M. Hassanizadeh, *Non-equilibrium interphase heat and mass transfer during two-phase flow in porous media—Theoretical considerations and modeling.* Advances in Water Resources, 2009. **32**(12): p. 1756-1766.

170. Puiroux, N., M. Prat, and M. Quintard, *Non-equilibrium theories for macroscale heat transfer: ablative composite layer systems.* International Journal of Thermal Sciences, 2004. **43**(6): p. 541-554.

171. Quintard, M., et al., *Nonlinear, multicomponent, mass transport in porous media.* Chemical Engineering Science, 2006. **61**(8): p. 2643-2669.

172. Stumm, W. and J.J. Morgan, *Aquatic chemistry. Chemical equilibria and rates in natural waters*. 3 ed. 1996, New York: Wiley-Interscience.

173. Morel, F.M.M., *Principles of aquatic chemistry*. 1983, New York: Wiley Interscience.

174. Lützenkirchen, J., *Description des interactions aux interfaces liquide-solide à l'aide des modèles de complexation et de précipitation de surface.* 1996, Université Louis Pasteur, Strasbourg. p. 351.

175. Lützenkirchen, J. and P. Behra, *A new approach for modelling potential effects in cation adsorption onto binary (hydr) oxides.* Journal of Contaminant Hydrology, 1997. **26**(1-4): p. 257-268.

176. Sigg, L., P. Behra, and W. Stumm, *Chimie des milieux aquatiques*. 3 ed. 2000, Paris: DUNOD.

177. Dzombak, D.A. and F.M.M. Morel, *Surface complexation modelling: Hydrous Ferric Oxide*. 1990, New-York: Wiley-Intersciences.

178. Lutzenkirchen, J. and P. Behra, *On the surface precipitation model for cation sorption at the (hydr)oxide water interface.* Aquatic Geochemistry, 1995. **1**(4): p. 375-397.

179. Morel, F. and J. Morgan, *A numerical method for computing equilibria in aqueous chemical systems.* Environmental Science and Technology, 1972. **6**(1): p. 58-67.

180. Westall, J.C., *MICROQL: a chemical equilibrium program in BASIC. Computation of adsorption equilibria in BASIC.*, S.F.I.o.T. EAWAG, Editor. 1979: Dübandorf. p. 42.

181. Steefel, C.I. and K.T.B. MacQuarrie, *Approaches to modeling of reactive transport in porous media.* Reviews in Mineralogy, 1996. **34**: p. 82-129.

182. Lichtner, P.C., R.T. Pabalan, and C.I. Steefel. *Model calculations of porosity reduction resulting from cement-tuff diffusive interaction*. in *Symposiumon Scientific Basis for Nuclear Waste Management XXI*. 1997. Davos, Switz.

183. Hunter, K.S., Y. Wang, and P. Van Cappellen, *Kinetic modeling of microbially-driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry.* Journal of Hydrology, 1998. **209**(1–4): p. 53-80.

184. Nilsen, V., J.A. Wyller, and A. Heistad, *Efficient incorporation of microbial metabolic lag in subsurface transport modeling.* Water Resources Research, 2012. **48**(9).

185. Mayer, K.U., E.O. Frind, and D.W. Blowes, *Multicomponent reactive transport modeling in variably saturated porous media using a generalized formulation for kinetically controlled reactions.* Water Resources Research, 2002. **38**(9): p. 13-1-13-21.

186. Parkhurst, D.L. and C.A.J. Appelo, *User's guide to PHREEQC (version 2)- A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations.*, Water-Resour.Invest., Editor. 1999: *Denver, CO, USA.* p. 312.

187. Hoteit, H., et al., *Three-dimensional modeling of mass transfer in porous media using the mixed hybrid finite elements and the random-walk methods.* Mathematical Geology, 2002. **34**(4): p. 435-456.

188. Younes, A., P. Ackerer, and F. Lehmann, *A new mass lumping scheme for the mixed hybrid finite element method.* International Journal for Numerical Methods in Engineering, 2006. **67**(1): p. 89-107.

189. Younes, A., et al., *A New Formulation of the Mixed Finite Element Method for Solving Elliptic and Parabolic PDE with Triangular Elements.* Journal of Computational Physics, 1999. **149**(1): p. 148-167.

190. Lefèvre, F., M. Sardin, and D. Schweich, *Migration of strontium in clayey and calcareous sandy soil: Precipitation and ion exchange.* Journal of Contaminant Hydrology, 1993. **13**(1–4): p. 215-229.

191. Barry, D.A., et al., *Comparison of split-operator methods for solving coupled chemical non-equilibrium reaction/groundwater transport models.* Mathematics and Computers in Simulation, 2000. **53**(1-2): p. 113-127.

192. Barry, D.A., et al., *Analysis of split operator methods for nonlinear and multispecies groundwater chemical transport models.* Mathematics and Computers in Simulation, 1997. **43**(3-6): p. 331-341.

193. Barry, D.A., C.T. Miller, and P.J. Culligan-Hensley, *Temporal discretisation errors in non-iterative split-operator approaches to solving chemical reaction/groundwater transport models.* Journal of Contaminant Hydrology, 1996. **22**(1-2): p. 1-17.

194. Colombo, R. and A. Corli, *On the Operator Splitting Method: Nonlinear Balance Laws and a Generalization of Trotter-Kato Formulas*, in *Hyperbolic Problems and Regularity Questions*, M. Padula and L. Zanghirati, Editors. 2007, Birkhäuser Basel. p. 91-100.

195. Kaluarachchi, J.J. and J. Morshed, *Critical assessment of the operator-splitting technique in solving the advection--dispersion--reaction equation: 1. First-order reaction.* Advances in Water Resources, 1995. **18**(2): p. 89-100.

196. Kanney, J.F., C.T. Miller, and C.T. Kelley, *Convergence of iterative split-operator approaches for approximating nonlinear reactive transport problems.* Advances in Water Resources, 2003. **26**(3): p. 247-261.

197. Lanser, D. and J.G. Verwer, *Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling.* Journal of Computational and Applied Mathematics, 1999. **111**(1-2): p. 201-216.

198. Morshed, J. and J.J. Kaluarachchi, *Critical assessment of the operator-splitting technique in solving the advection--dispersion--reaction equation: 2. Monod kinetics and coupled transport.* Advances in Water Resources, 1995. **18**(2): p. 101-110.

199. Simpson, M.J. and K.A. Landman, *Theoretical analysis and physical interpretation of temporal truncation errors in operator split algorithms.* Mathematics and Computers in Simulation, 2008. **77**(1): p. 9-21.

200. Wigley, T.M.L., *WATSPEC: A computer program for determining the equilibrium speciation of aqueous solutions.*, B.G.R.G. Tech.Bull., Editor. 1977. p. 49.

201. Brassard, P. and P. Bodurtha, *A feasible set for chemical speciation problems.* Computers & Geosciences, 2000. **26**(3): p. 277-291.

202. Erhel, J. and S. Sabit, *Analysis of a global reactive transport model and results for the MoMaS benchmark.* Mathematics and Computers in Simulation, 2017. **137**: p. 286-298.

203. Machat, H. and J. Carrayrou, *Comparison of linear solvers for equilibrium geochemistry computations.* Computational Geosciences, 2017. **21**(1): p. 131-150.

204. Golub, H.V.v.L., C.F., *Matrix computations*. 3rd ed. 1996, Baltimore: The Johns Hopkins University Press.

205. Kincaid, D. and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing*. 3 ed. 2002: American Mathematical Society.

206. Quarteroni, A., Sacco, R., Saleri, F., *Numerical Mathematics*. 2 ed. Texts in Applied Mathematics, ed. J.E. Marsden, L. Sirovich, and S.S. Antman. 2007, Heidelberg: Springer.

207. Marinoni, M., Carrayrou, J., Lucas, Y., Ackerer, P., *Thermodynamic equilibrium solutions through a modified Newton Raphson method.* AIChE Journal, 2016.

208. Knight, P., D. Ruiz, and B. Ucar, *A symmetry preserving algorithm for matrix scaling.* SIAM Journal on Matrix Analysis and Applications, 2014. **35**(3): p. 931-955.

209. Bahr, J.M. and J. Rubin, *Direct comparison of kinetic and local equilibrium formulation for solute transport affected by surface reactions.* Water Resources Research, 1987. **23**(3): p. 438-452.

210. Jennings, A.A. and D.J. Kirkner, *Instantaneous equilibrium approximation analysis.* Journal of Hydraulic Engineering, 1984. **110**(12): p. 1700-1717.

211. Rubin, J., *Transport of reacting solutes in porous media: relation between mathematical nature of problem formulation and chemical nature of reactions.* Water Resources Research, 1983. **19**(5): p. 1231-1252.

212. Valocchi, A.J., *Validity of the local equilibrium assumption for modeling sorbing solute transport through homogeneous soils.* Water Resour.Res., 1985. **21**: p. 808-820.

213. Bahr, J.M. and J. Rubin, *Direct comparison of kinetic and local equilibrium formulations for solute transport affected by surface reactions.* Water Resources Research, 1987. **23**(3): p. 438-452.

214. Sandu, A., et al., *Benchmarking stiff ode solvers for atmospheric chemistry problems II: Rosenbrock solvers.* Atmospheric Environment, 1997. **31**(20): p. 3459-3472.

215. Sandu, A., et al., *Benchmarking stiff ode solvers for atmospheric chemistry problems-I. implicit vs explicit.* Atmospheric Environment, 1997. **31**(19): p. 3151-3166.

216. Verwer, J.G., et al., *A comparison of stiff ODE solvers for atmospheric chemistry problems.* Atmospheric Environment, 1996. **30**(1): p. 49-58.

217. Richardson, L.F., *The approximate arithmetical solution by finite difference of physical problems involving differential equations, with an application to the stress in a masonry dam.* Philos. Trans. Roy. Soc. London, 1910. **210**(A): p. 307-357.

218. Richardson, L.F., *The deferred approach to the limit. I: single lattice.* Philos. Trans. Roy. Soc. London, 1927. **226**(A): p. 299-349.

219. Hairer, E., Nörsett, S. P. and Wanner, G, *Solving Ordinary Equations I, Nonstiff problems*. 2 ed. 2000, Berlin: Springer-Verlag.

220. Bastian, P. and S. Lang, *Couplex Benchmark Computations Obtained with the Software Toolbox UG.* Computational Geosciences, 2004. **8**(2): p. 125-147.

221. Bernard-Michel, G., et al., *The Andra Couplex 1 Test Case: Comparisons Between Finite-Element, Mixed Hybrid Finite Element and Finite Volume Element Discretizations.* Computational Geosciences, 2004. **8**(2): p. 187-201.

222. Bourgeat, A., et al., *The COUPLEX Test Cases: Nuclear Waste Disposal Simulation.* Computational Geosciences, 2004. **8**(2): p. 83-98.

223. Bourgeat, A.P. and M. Kern, *Special Issue on Simulation of Transport around a Nuclear Waste Disposal Site: The COUPLEX Test Cases.* Computational Geosciences, 2004. **8**(2): p. 81-82.

224. Chénier, E., R. Eymard, and X. Nicolas, *A Finite Volume Scheme for the Transport of Radionucleides in Porous Media.* Computational Geosciences, 2004. **8**(2): p. 163-172.

225. Del Pino, S. and O. Pironneau, *Asymptotic Analysis and Layer Decomposition for the Couplex Exercise.* Computational Geosciences, 2004. **8**(2): p. 149-162.

226. Lanteri, S. and C. Raffourt, *Strategies for Reducing Computing Time of Nuclear Waste Management Simulations Using the PORFLOW ™ Software.* Computational Geosciences, 2004. **8**(2): p. 203-215.

227. Trujillo, D., *Mixed Primal–Dual Method for Nuclear Waste Disposal Far Field Simulation.* Computational Geosciences, 2004. **8**(2): p. 173-185.

228. Oreskes, N., K. Shrader-Frechette, and K. Belitz, *Verification, validation, and confirmation of numerical models in the earth sciences.* Science, 1994. **263**(5147): p. 641-646.

229. O'Connor, G.A. and P.J. Wierenga, *The persistence of 2,4,5-T in greenhouse lysimeter studies.* Soil Science Society of America Journal, 1973. **37**: p. 398-400.

230. Gamerdinger, A.P., R.J. Wagenet, and M.T. van Genuchten, *Application of two-site/two-region models for studying simultaneous nonequilibrium transport and degradation of pesticides.* Soil Science Society of America Journal, 1990. **54**(4): p. 957-963.

231. Brusseau, M.L., R.E. Jessup, and P.S.C. Rao, *Modeling the transport of solutes influenced by multiprocess nonequilibrium.* Water Resources Research, 1989. **25**(9): p. 1971-1988.

232. Nitzsche, O., G. Meinrath, and B. Merkel, *Database uncertainty as a limiting factor in reactive transport prognosis.* Journal of Contaminant Hydrology, 2000. **44**(3-4): p. 223-237.

233. Leij, F.J., N. Toride, and M.T. van Genuchten, *Analytical solutions for non-equilibrium solute transport in three-dimensional porous media.* Journal of Hydrology, 1993. **151**(2-4): p. 193-228.

234. Lu, X., Y. Sun, and J.N. Petersen, *Analytical Solutions of TCE Transport with Convergent Reactions.* Transport in Porous Media, 2003. **51**(2): p. 211-225.

235. Selim, H.M. and R.S. Mansell, *Analytical solution of the equation for transport of reactive solutes through soils.* Water Resources Research, 1976. **12**(3): p. 528-532.

236. Severino, G. and P. Indelman, *Analytical solutions for reactive transport under an infiltration-redistribution cycle.* Journal of Contaminant Hydrology, 2004. **70**(1-2): p. 89-115.

237. Srinivasan, V. and T.P. Clement, *Analytical solutions for sequentially coupled one-dimensional reactive transport problems - Part II: Special cases, implementation and testing.* Advances in Water Resources, 2008. **31**(2): p. 219-232.

238. Srinivasan, V. and T.P. Clement, *Analytical solutions for sequentially coupled one-dimensional reactive transport problems - Part I: Mathematical derivations.* Advances in Water Resources, 2008. **31**(2): p. 203-218.

239. Srivastava, R., P.K. Sharma, and M.L. Brusseau, *Reactive solute transport in macroscopically homogeneous porous media: analytical solutions for the temporal moments.* Journal of Contaminant Hydrology, 2004. **69**(1-2): p. 27-43.

240. Sun, Y. and T.A. Buscheck, *Analytical solutions for reactive transport of N-member radionuclide chains in a single fracture.* Journal of Contaminant Hydrology, 2003. **62-63**: p. 695-712.

241. Sun, Y., et al., *An Analytical Solution of Tetrachloroethylene Transport and Biodegradation.* Transport in Porous Media, 2004. **55**(3): p. 301-308.

242. Sun, Y., J.N. Petersen, and T.P. Clement, *Analytical solutions for multiple species reactive transport in multiple dimensions.* Journal of Contaminant Hydrology, 1999. **35**(4): p. 429-440.

243. Tartakovsky, D.M., *An analytical solution for two-dimensional contaminant transport during groundwater extraction.* Journal of Contaminant Hydrology, 2000. **42**(2-4): p. 273-283.

244. Toride, N., F.J. Leij, and M.T. Van Genuchten, *A comprehensive set of analytical solutions for nonequilibrium solute transport with first-order decay and zero-order production.* Water Resources Research, 1993. **29**(7): p. 2167-2182.

245. Van Der Zee, S.E.A.T., *Analytical traveling wave solutions for transport with nonlinear and nonequilibrium adsorption.* Water Resources Research, 1990. **26**(10): p. 2563-2578.

246. van Genuchten, M.T., *Analytical solutions for chemical transport with simultaneous adsorption, zero-order production and first-order decay.* Journal of Hydrology, 1981. **49**(3-4): p. 213-233.

247. van Genuchten, M.T. and P.J. Wierenga, *Mass transfer studies in sorbing porous media - 1. Analytical solutions.* Soil Science Society of America Journal, 1976. **40**(4): p. 473-480.

248. Loyaux-Lawniczak, S., F. Lehmann, and P. Ackerer, *Acid/base front propagation in saturated porous media: 2D laboratory experiments and modeling.* Journal of Contaminant Hydrology, 2012. **138–139**(0): p. 15-21.

249. Lu, J., C. Beaucaire, and E. Tertre, *Predictive Model for Migration of Metallic Cations in Natural Sediments.* Procedia Earth and Planetary Science, 2013. **7**: p. 529-532.

250. Bürgisser, *Transportverhalten von nicht-linear sorbierenden Stoffen in chromatographischen Säulen am Beispiel der Adsorption an Cristobalit*. 1994, ETH Zürich: Zürich. p. 145.

251. Thaysen, E.M., et al., *Effect of dissolved H2SO4 on the interaction between CO2-rich brine solutions and limestone, sandstone and marl.* Chemical Geology, 2017. **450**: p. 31-43.

252. Prommer, H., D.A. Barry, and C. Zheng, *MODFLOW/MT3DMS-Based Reactive Multicomponent Transport Modeling.* Ground Water, 2003. **41**(2): p. 247-257.

253. Trotignon, L., et al., *Intercomparison between TRIO-EF and IMPACT codes with reference to experimental strontium migration data.* Journal of Contaminant Hydrology, 1997. **26**(1-4): p. 279-289.

254. de Dieuleveult, C., J. Erhel, and M. Kern, *A global strategy for solving reactive transport equations.* Journal of Computational Physics, 2009. **228**(17): p. 6395-6410.

255. Sabit, S., *Les méthodes numériques de transport réactif*. 2014, Rennes 1. p. 137.

256. CEA, *CASTEM2000 user's manual. Tech. Rep.* . 1999, CEA (France).

257. Lucille, P.L., A. Burnol, and P. Ollar, *Chemtrap: a hydrogeochemical model for reactive transport in porous media.* Hydrological Processes, 2000. **14**(13): p. 2261-2277.

258. van der Lee, J., et al., *Presentation and application of the reactive transport code HYTEC*, in *Developments in Water Science, Computational Methods in Water Resources, Proceedings of the XIVth International Conference on Computational Methods in Water Resources (CMWR XIV)*, S.M. Hassanizadeh, Editor. 2002, Elsevier. p. 599-606.

259. Meeussen, J.C.L., *ORCHESTRA: An Object-Oriented Framework for Implementing Chemical Equilibrium Models.* Environmental Science & Technology, 2003. **37**(6): p. 1175-1182.

260. Roose, T., A.C. Fowler, and P.R. Darrah, *A mathematical model of plant nutrient uptake.* Journal of Mathematical Biology, 2001. **42**(4): p. 347-360.

261. Steefel, C.I. and K. Maher, *Fluid-rock interaction: A reactive transport approach.* Reviews in mineralogy and geochemistry, 2009. **70**(1): p. 485-532.

262. Steefel, C., *Crunch – user's guide*. 2006, USA: Lawrence Berkeley Laboratory.

263. Arora, B., et al., *A reactive transport benchmark on heavy metal cycling in lake sediments.* Computational Geosciences, 2014.

264. Xu, T., et al., *TOUGHREACT Version 2.0: A simulator for subsurface reactive transport under non-isothermal multiphase flow conditions.* Computers & Geosciences, 2011. **37**(6): p. 763-774.

265. Steefel, C., Molins, S., *CRUNCHFLOW: Software for Modeling Multicomponent Reactive Flow and Transport. User's manual*. 2016, USA: Lawrence Berkeley Laboratory.

266. Wanner, C., et al., *Benchmarking the simulation of Cr isotope fractionation.* Computational Geosciences, 2015. **19**(3): p. 497-521.

267. Lichtner, P.C., *FLOTRAN users manual: two-phase nonsothermal coupled thermal-hydrologic-chemical (THC) reactive flow and transport code version 2.* 2007, Los Alamos National Laboratory: New Mexico, USA.

268. Saaltink, M.W., J. Carrera, and C. Ayora. *On the numerical formulation of reactive transport problems*. in *11th International Conference on Computational Methods in Water Resources, CMWR'96*. 1996. Cancun, Mex.

269. Vag, J.E., W. Hong, and H.K. Dahle, *Eulerian-Lagrangian localized adjoint methods for systems of nonlinear advective-diffusive-reactive transport equations.* Advances in Water Resources, 1996. **19**(5): p. 297-315.

270. Saaltink, M.W., C. Ayora, and J. Carrera, *A mathematical formulation for reactive transport that eliminates mineral concentrations.* Water Resources Research, 1998. **34**(7): p. 1649-1656.

271. Kanney, J.F., C.T. Miller, and D.A. Barry, *Comparison of fully coupled approaches for approximating nonlinear transport and reaction problems.* Advances in Water Resources, 2003. **26**(4): p. 353-372.

272. Carrayrou, J., *First comparative analysis of benchmark's results*, in *International workshop on reactive transport*. 2008: Strasbourg.

273. Lehmann, F. and P. Ackerer. *Inverse problem for one-dimensional subsurface flow in unsaturated porous media*.

274. Siegel, P., A.P. Blaschke, and P. Ackerer. *Inverse problem applied to groundwater flow and transport equations using a downscaling parameterization*.

275. Stockel, M.E., R. Mose, and P. Ackerer. *Application of the sentinel method in a groundwater transport model*.

276. Allison JD, B.D., Novo-Gradac KJ., *MINTEQA2/PRODEFA2. A Geochemical Assessment Model for Environmental Systems: Version 3.0 User's Manual.* 1990, Environmental Research Laboratory, Office of Research and Development, USEPA, Athens, CA.

277. Ludwig, C., *GRFIT: A Program, for Solving Speciation Problems, Evaluation of Equilibrium Constants, Concentrations and Their Physical Parameters.* 1992, Switzerland: The University of Berne, 1992.

278. Bajracharya, K. and D.A. Barry, *MCMFIT: Efficient optimal fitting of a generalized nonlinear advection-dispersion model to experimental data.* Computers and Geosciences, 1995. **21**(1): p. 61-76.

279. McGrail, B.P., *Inverse reactive transport simulator (INVERTS): An inverse model for contaminant transport with nonlinear adsorption and source terms.* Environmental Modelling and Software, 2001. **16**(8): p. 711-723.

280. Schweich, D. and M. Sardin, *Adsorption, partition, ion exchange and chemical reaction in batch reactors or in columns - A review.* Journal of Hydrology, 1981. **50**: p. 1-33.

281. Schweich, D. and M. Sardin, *Transient ion exchange and solubilization of limestone in an oil field sandstone: experimental and theoretical wavefront analysis.* AIChE Journal, 1985. **31**(11): p. 1882-1890.

282. Schweich, D., M. Sardin, and M. Jauzein, *From water movement to solute transport: a problem of physico-chemistry.* Bulletin - Societe Geologique de France, 1988. **4**(5): p. 879-886.

283. Schweich, D., M. Sardin, and M. Jauzein, *Properties of concentration waves in presence of nonlinear sorption, precipitation/dissolution, and homogeneous reactions. 1. Fundamentals.* Water Resources Research, 1993. **29**(3): p. 723-733.

284. Bueno, M., *Etude dynamique des processus de sorption-désorption du tributylétain sur un milieu poreux d'origine naturelle*. 1999, Université de Pau et des Pays de l'Adour.

285. Loyaux-Lawniczak, S., et al., *Trapping of Cr by formation of ferrihydrite during the reduction of chromate ions by Fe(II)-Fe(III) hydroxysalt green rusts.* Environmental Science & Technology, 2000. **34**(3): p. 438-443.

286. Loyaux-Lawniczak, S., et al., *The reduction of chromate ions by Fe(II) layered hydroxides.* Hydrology and Earth System Sciences, 1999. **3**(4): p. 593-599.

287. Carrayrou, J., et al., *Modeling laboratory scale experiment (TRACE and SPECY) of Fe-Cr redox reaction along with precipitation and porosity change.*, in *MAMERN VII*. 2017: Oujda, Maroco.

288. Applin, K.R. and A.C. Lasaga, *The determination of SO42-, NaSO4-, and MgSO40 tracer diffusion coefficients and their application to diagenetic flux calculations.* Geochimica et Cosmochimica Acta, 1984. **48**(10): p. 2151-2162.

289. Ben-Yaakov, S., *Diffusion of sea water ions—I. Diffusion of sea water into a dilute solution.* Geochimica et Cosmochimica Acta, 1972. **36**(12): p. 1395-1406.

290. Ben-Yaakov, S., *Discussion—diffusion of seawater ions: significance and consequences of cross coupling effects.* Am. J. Sci., 1981. **281**: p. 974-980.

291. Boudreau, B.P., F.J.R. Meysman, and J.J. Middelburg, *Multicomponent ionic diffusion in porewaters: Coulombic effects revisited.* Earth and Planetary Science Letters, 2004. **222**(2): p. 653-666.

292. Lasaga, A.C., *Influence of diffusion coupling on diagenetic concentration profiles.* Am. J. Sci, 1981. **281**: p. 553-575.

293. Lasaga, A.C., *The treatment of multicomponent diffusion and ion pairs in diagenetic fluxes.* Am. J. Sci, 1979. **279**: p. 324-346.

294. Rasouli, P., et al., *Benchmarks for multicomponent diffusion and electrochemical migration.* Computational Geosciences, 2015. **19**(3): p. 523-533.

295. Steefel, C.I., et al., *Reactive transport codes for subsurface environmental simulation.* Computational Geosciences, 2015. **19**(3): p. 445-478.

296. Keyes, D.E., et al., *Multiphysics simulations: Challenges and opportunities.* International Journal of High Performance Computing Applications, 2013. **27**(1): p. 4-83.

297. Simmons, C.T., T.R. Fenstemaker, and J.M. Sharp Jr, *Variable-density groundwater flow and solute transport in heterogeneous porous media: approaches, resolutions and future challenges.* Journal of Contaminant Hydrology, 2001. **52**(1–4): p. 245-275.

298. Lasaga, A.C. and R.A. Berner, *Fundamental aspects of quantitative models for geochemical cycles.* Chemical Geology, 1998. **145**(3–4): p. 161-175.

299. Younes, A., et al., *The mixed finite element method with one unknown per element.* Comptes Rendus de l'Academie des Sciences - Series I: Mathematics, 1999. **328**(7): p. 623-626.

300. Hoteit, H., et al., *The maximum principle violations of the mixed-hybrid finite-element method applied to diffusion equations.* International Journal for Numerical Methods in Engineering, 2002. **55**(12): p. 1373-1390.

301. Ackerer, P., A. Younes, and R. Mose, *Modeling variable density flow and solute transport in porous medium: 1. Numerical model and verification.* Transport in Porous Media, 1999. **35**(3): p. 345-373.

302. Ackerer, P., et al., *On modelling of density driven flow.* IAHS-AISH Publication, 2000(265): p. 377-384.

303. Diaw, E.B., F. Lehmann, and P. Ackerer, *One-dimensional simulation of solute transfer in saturated-unsaturated porous media using the discontinuous finite elements method.* Journal of Contaminant Hydrology, 2001. **51**(3-4): p. 197-213.

304. Diaw, E.H.B., F. Lehmann, and P. Ackerer, *Modeling a non-conservative solute transfer in unsaturated porous media.* Comptes Rendus de l'Academie de Sciences - Serie IIa: Sciences de la Terre et des Planetes, 2001. **333**(2): p. 129-132.

305. Lehmann, F. and P. Ackerer, *Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media.* Transport in Porous Media, 1998. **31**(1-3): p. 275-292.

Annexes

Annexe 1. Modélisation du transport réactif en milieu poreux : schéma itératif associé à une combinaison d'éléments finis discontinus et mixtes-hybrides

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
C. R. Mecanique 331 (2003) 211–216
COMPTES RENDUS MECANIQUE

Modélisation du transport réactif en milieu poreux : schéma itératif associé à une combinaison d'éléments finis discontinus et mixtes-hybrides

Modelling reactive transport in porous media: iterative scheme and combination of discontinuous and mixed-hybrid finite elements

Jérôme Carrayrou [a], Robert Mosé [a,b], Philippe Behra [a,c]

[a] *Institut de mécanique des fluides et des solides de l'Université Louis Pasteur, UMR 7507, Université Louis Pasteur – CNRS, 2, rue Boussingault, 67000 Strasbourg, France*
[b] *École nationale du génie de l'eau et de l'environnement de Strasbourg, 1, quai Koch, 67000 Strasbourg, France*
[c] *École nationale supérieure des ingénieurs en arts chimiques et technologiques, Laboratoire de chimie agro-industrielle, UMR 1010 INRA/INP-ENSIACET, 118, route de Narbonne, 31077 Toulouse cedex 4, France*

Reçu le 17 septembre 2001 ; accepté après révision le 14 janvier 2003

Présenté par Michel Combarnous

Résumé

Dans le cadre d'une approche par séparation d'opérateurs, le schéma itératif (I) représente une bonne méthode pour la résolution du transport de solutés réactifs en milieu poreux. La combinaison d'éléments finis discontinus et mixtes-hybrides permet la résolution efficace du transport en milieu poreux, mais sa mise en œuvre dans un schéma I conduit à des difficultés numériques importantes. Nous présentons ici comment associer ces deux méthodes afin de simuler le transport de solutés réactifs. La modélisation d'une expérience en colonne permet de tester cette nouvelle méthode. ***Pour citer cet article : J. Carrayrou et al., C. R. Mecanique 331 (2003).***

Abstract

The sequential iterative approach (SIA) scheme is the most efficient method for modelling reactive transport in porous media with the operator-splitting approach. A combination of finite discontinuous and finite mixed-hybrid elements is a powerful method for solving solute transport in porous media, but the use of this method for SIA scheme induces numerical difficulties. In this paper, a new method is developed to solve reactive transport by using both the SIA scheme and a combination of finite discontinuous and finite mixed elements. The proposed method is tested by modelling a column experiment. ***To cite this article: J. Carrayrou et al., C. R. Mecanique 331 (2003).***

Adresse e-mail : carrayro@imfs.u-strasbg.fr (J. Carrayrou).

1631-0721/03/$ – see front matter
doi:10.1016/S1631-0721(03)00040-8

Mots-clés : Milieux granulaires ; Sols ; Milieux poreux ; Transferts thermiques ; Milieux réactifs ; Combustion ; Transport réactif ; Séparation d'opérateurs ; Schéma itératif ; Éléments finis discontinus ; Éléments finis mixtes-hybrides

Keywords: Granular media; Soils; Porous media; Heat transfer; Reactive media; Combustion; Reactive transport; Operator splitting; Iterative scheme; Finite discontinuous elements; Mixed-hybrid finite elements

Abridged English version

Modelling reactive transport in porous media needs the resolution of both a transport and a chemistry operator. In an operator-splitting (OS) approach, each operator is solved independently of the other one. The OS approach is very attractive because it is computationally less expensive than solving the global system and it allows the use of highly specific method for solving each operator. Nevertheless, it induces intrinsic OS errors due to the separation of the operators. In this paper, we discuss a way to associate a very efficient method for solving the transport operator with an OS approach that minimizes the OS errors. Under an instantaneous equilibrium formulation, it has been well reported [2,7] that the Standard Iterative (SI) scheme is the OS method that induces the minimum of OS errors, but it needs an implicit time formulation for the transport operator. On the other hand, the Standard Non Iterative (SNI) scheme induces more OS errors but the time formulation of the transport operator is free. The method (EFDMH) used for solving transport operator [3] is based on the separation of the two parts of the transport equation. The advective (the dispersive) part is solved with a discontinuous finite elements method with an explicit time formulation (mixed hybrid finite elements method with an implicit time formulation, respectively).

We associated EFDMH and SI scheme in a two steps algorithm. First step: the advective part of the transport equation was solved with discontinuous finite elements and the chemistry operator was then solved and associated to a SNI scheme. Second step: the dispersive part was solved with mixed hybrid finite elements and the chemistry operator was solved and associated to a SI scheme.

These associations were tested and compared. A reactive transport experiment (Table 1) from Lefèvre et al. [9] was solved by three ways: (i) The transport operator solved by finite differences with an implicit time formulation and OS is done by a SI scheme; (ii) EFDMH associated with the SNI scheme; (iii) EFDMH and SI scheme associated as written above. A reference solution was obtained by the three ways, if the mesh and the time step were sufficiently small. With larger mesh and time step, we showed (Fig. 1) that the ways (i) and (iii) were the less and the more efficient, respectively. This shows the importance of associating the best method for solving each operator and the best OS scheme for solving reactive transport problems.

1. Introduction

La compréhension des phénomènes de transport de solutés réactifs ou transport réactif en milieu poreux et leur simulation dans un but prédictif constituent une étape incontournable pour une gestion durable de la qualité des aquifères ou du stockage des déchets. Les modèles numériques de transport de solutés réactifs en milieu poreux représentent à ce titre des outils privilégiés.

Actuellement, ces modèles portent essentiellement sur le transport réactif dans des domaines mono-directionnels, éventuellement bi-dimensionnels, et homogènes tant d'un point de vue (géo)chimique qu'hydrodynamique [1]. Pour appliquer ces modèles à des cas réels, il s'avère indispensable de prendre en compte des domaines bi- et tri-dimensionnels hétérogènes. Dans le cadre d'une approche par séparation d'opérateurs [2], il est nécessaire d'utiliser des méthodes spécifiques de résolution de l'opérateur de transport permettant la prise en compte performante de ces domaines 2D ou 3D. Or leur mise en œuvre dans le transport réactif se heurte à de nombreux obstacles numériques. Dans cet article, nous présenterons comment surmonter ces difficultés et utiliser une combinaison d'éléments finis discontinus et d'éléments finis mixtes hybrides (EFDMH) pour résoudre l'opé

rateur de transport [3]. L'utilisation de cette combinaison EFDMII permet d'obtenir une solution compatible avec les contraintes chimiques, absence de concentrations négatives, faible diffusion numérique, quelles que soient les conditions hydrodynamiques. Il est cependant nécessaire d'adapter les techniques de séparation d'opérateurs pour obtenir les meilleurs résultats.

Si la vitesse de l'écoulement est suffisamment lente par rapport aux temps caractéristiques des réactions [4], les phénomènes chimiques peuvent être supposés à l'équilibre thermodynamique. La représentation des systèmes chimiques par tableaux des équilibres [5] permet de ne résoudre l'opérateur de transport que pour les concentrations totales en composants [6,7], en supposant le tenseur de dispersion $\mathbf{D}$ identique pour chaque forme chimique d'un élément donné [8] :

$$\frac{\partial(\omega Td_j + \rho_S Tf_j)}{\partial t} = \nabla\cdot\left[\mathbf{D}\cdot\nabla(Td_j)\right] \quad U\cdot\nabla(Td_j) \tag{1}$$

avec ω la porosité, ρ_S la masse volumique de la matrice solide, U la vitesse de Darcy, et pour le composant j (soluté), Td_j la concentration totale en soluté dissous et Tf_j la concentration totale en soluté fixé. Par soucis de simplification, ω et ρ_S seront supposés constants. La répartition des espèces entre les phase aqueuse et solide est calculée par résolution du système algébrique non linéaire formé par les lois d'action de masse et les équations de conservation de la matière.

2. Méthodes

L'approche par séparation d'opérateurs permet de résoudre séparément les opérateurs de transport et de chimie et d'utiliser les méthodes les mieux adaptées pour chaque opérateur. Deux schémas de séparation sont alors envisageables [7].

2.1. Schéma Non Itératif (NI)

Le schéma NI entre les pas de temps n et $n+1$ consiste en une étape de transport non réactif (2) suivie du calcul des équilibres thermodynamiques (3) :

$$\omega\frac{Td_j^* - Td_j^n}{\Delta t} \quad \nabla\cdot\left[\mathbf{D}\cdot\nabla(Td_j)\right] - U\cdot\nabla(Td_j) \tag{2}$$

où Td_j^* est la solution de l'opérateur de transport au pas de temps $n+1$:

$$\omega Td_j^{n+1} = f_d\left(\omega Td_j^* + \rho_S Tf_j^n\right) \quad \text{et} \quad \rho_S Tf_j^{n+1} = f_f\left(\omega Td_j^* + \rho_S Tf_j^n\right) \tag{3}$$

f_d et f_f représentent les systèmes algébriques non linéaires formés par les lois d'action de masse et les équations de conservation de la matière donnant les concentrations totales en soluté dissous et fixé des composants en fonction des concentrations totales. Td_j^{n+1} et Tf_j^{n+1} sont les solutions du problème de transport réactif au pas de temps $n+1$.

2.2. Schéma Itératif (I)

Le schéma I, à l'itération $k+1$ entre les pas de temps n et $n+1$, se décompose en une étape de transport réactif (4), dans laquelle le caractère réactif est introduit sous la forme d'un terme puits-source, et en une étape d'actualisation du terme puits-source par calcul des équilibres (5) :

$$\omega\frac{Td_j^{n+1,k+1,*} - Td_j^n}{\Delta t} - \nabla\cdot\left[\mathbf{D}\cdot\nabla(Td_j)\right] - U\cdot\nabla(Td_j) - \rho_S\frac{Tf_j^{n+1,k} - Tf_j^n}{\Delta t} \tag{4}$$

où $Td_j^{n+1,k+1,*}$ est la solution de l'opérateur de transport au pas de temps $n-1$ à l'itération $k+1$:

$$\omega Td_j^{n+1,k+1} \quad f_d\left(\omega Td_j^{n+1,k+1,*} + \rho_S Tf_j^{n+1,k}\right) \quad \text{et} \quad \rho_S Tf_j^{n+1,k-1} \quad f_f\left(\omega Td_j^{n+1,k+1,*} - \rho_S Tf_j^{n+1,k}\right) \tag{5}$$

$Td_j^{n+1,k-1}$ et $Tf_j^{n+1,k+1}$ représentent les solutions du transport réactif au pas de temps $n+1$ et à l'itération $k+1$. La solution exacte est approchée par la méthode du point fixe.

Le schéma I est plus précis que le schéma NI, mais il requiert une formulation implicite en temps pour la résolution de l'opérateur de transport [7], alors que le schéma NI laisse toute liberté pour choisir la formulation temporelle la plus adaptée.

2.3. Résolution du transport : combinaison EFDMH

Lorsque le transport est fortement convectif, les méthodes numériques classiques induisent une forte diffusion numérique ainsi que des oscillations. Aussi est-il impératif de respecter certains critères pour assurer la stabilité du schéma. Une formulation explicite en temps nécessite le respect du critère de Courant, $Co = (U/\omega)\cdot(\Delta t/\Delta x) \leqslant 1$, où Co est le nombre de Courant. Le critère de Péclet impose $Pe = \Delta x/\alpha < 2$, quelle que soit la formulation temporelle choisie, où Pe est le nombre de Péclet de maille et α la dispersivité. L'utilisation d'une combinaison EFDMH permet de s'affranchir du critère de Péclet, de limiter la diffusion numérique et d'éviter les oscillations [3]. La partie convective de l'équation de transport est résolue par éléments finis discontinus. En respectant le critère de Courant, cette approche permet de résoudre l'opérateur de transport avec $0 < Pe < \infty$. La résolution temporelle adoptée est d'ordre deux en temps. La solution obtenue respecte le principe du maximum, ce qui garantit l'absence oscillation. La partie dispersive est résolue par éléments finis mixtes hybrides qui permettent d'obtenir un bilan de masse local exact ainsi qu'une prise en compte aisée d'un tenseur de dispersion $\mathbf{D}$ plein [3].

2.4. Association du Schéma Itératif et d'une combinaison EFDMH

La combinaison EFDMH est donc une méthode de résolution de l'opérateur de transport dont les performances répondent aux exigences de la modélisation du transport réactif. Cependant, seul un schéma de couplage NI semble applicable si l'opérateur de transport est résolu par cette méthode, la partie convective de l'équation de transport étant formulée explicitement en temps. Néanmoins, la partie dispersive est résolue par élément finis mixtes-hybrides, avec une formulation implicite en temps. Un schéma de résolution itératif peut donc être proposé. Un schéma NI permet la résolution de la partie convective de l'opérateur de transport par éléments finis discontinus (6) et le calcul des équilibres (7). Un processus itératif est ensuite mis en place entre (i) la résolution de la partie dispersive avec un terme puits-source réactif (8) par éléments finis mixtes-hybrides, et (ii) l'actualisation du terme puits-source réactif par calcul des équilibres (5) :

$$\omega\frac{Td_j^{k} - Td_j^{c}}{\Delta t} \quad -U\cdot\nabla\left(Td_j^{k}\right) \tag{6}$$

avec Td_j^{k} solution de la partie convective au pas de temps $n+1$;

$$\omega Td_j^{*,c} = f_d\left(\omega Td_j^{c} + \rho_S Tf_j^{n}\right) \quad \text{et} \quad \rho_S Tf_j^{*,c} = f_f\left(\omega Td_j^{c} + \rho_S Tf_j^{n}\right) \tag{7}$$

où $Td_j^{k,c}$ et $Tf_j^{k,*}$ sont les solutions de la partie convective-réactive au pas de temps $n+1$:

$$\omega\frac{Td_j^{c,k+1} - Td_j^{k,c}}{\Delta t} = \nabla\cdot\left[\mathbf{D}\cdot\nabla\left(Td_j^{n+1,k+1}\right)\right] - \rho_S\frac{Tf_j^{c,k+1} - Tf_j^{*,k}}{\Delta t} \tag{8}$$

$Td_j^{n+1,k-1}$ et $Tf_j^{n+1,k+1}$ représentent les solutions du problème de transport réactif au pas de temps $n+1$ et à l'itération $k+1$.

3. Discussion

Les performances des schémas précédents sont comparées à l'aide d'une expérience en colonne réalisée par Lefèvre et al. [9]. Il s'agit de la migration du strontium, sous forme Sr^{2+}, à travers un sol sableux argilo-calcaire (voir Tableau 1). La combinaison des phénomènes de précipitation, de dissolution et d'échange de cations qui ont

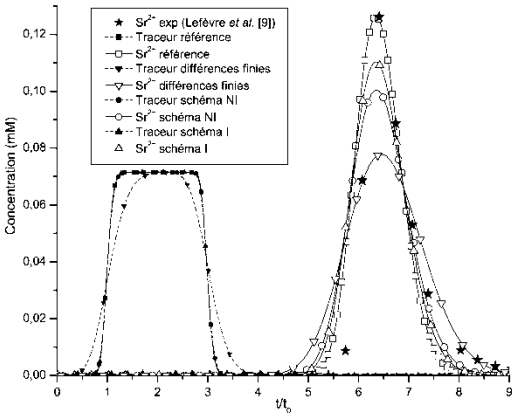


Fig. 1. Comparaison des courbes expérimentales d'élutions d'un traceur et du strontium [9] et de celles obtenues pour un calcul de référence (maillage fin : $\Delta t = 7.61 \times 10^{-4} t_0$ et $\Delta x = 10^{-4}$ m), des solutions calculées sur un maillage large ($\Delta t = 7.61 \times 10^{-3} t_0$ et $\Delta x = 10^{-3}$ m) pour un schéma NI et par un schéma I avec une combinaison EF/DMH, et pour un schéma I avec des différences finies implicites. (Conditions expérimentales d'après Lefèvre et al. [9], temps de séjour d'un traceur t_0, données des calculs cf. Tableau 1.)

Fig. 1. Comparison between experimental breakthrough curves of tracer and strontium [9], breakthrough curves for a reference calculation (fine mesh: Δt $7.61 \times 10^{-4} t_0$ and Δx 10^{-4} m), and breakthrough curves calculated over a large mesh (Δt $7.61 \times 10^{-3} t_0$ et Δx 10^{-3} m) by a non iterative scheme and a SIA scheme with a combination of discontinuous and mixed-hybrid finite elements and by a SIA scheme with implicit finite difference. (Experimental conditions after Lefèvre et al. [9], breakthrough time of a tracer t_0, computation parameters see Table 1.)

Tableau 1
Paramètres chimiques utilisés pour les calculs de l'expérience de sorption du strontium [9]

Table 1
Chemical parameters used for calculating the strontium sorption experiment [9]

Réaction	Constante d'équilibre pK
$H_2O = H^+ + OH^-$	14.0
$H_2CO_3^* \rightleftharpoons H^+ + HCO_3^-$	6.3
$HCO_3^- \rightleftharpoons H^+ + CO_3^{2-}$	10.3
$CaCO_3(s) = Ca^{2+} + CO_3^{2-}$	8.42
$SrCO_3(s) \rightleftharpoons Sr^{2+} + CO_3^{2-}$	9.03
$Ca_f^{2+} + Sr^{2+} = Ca^{2+} + Sr_f^{2+}$	0.021
Injection – Injection (Durée – Duration $= 2t_0$)	$[H_2CO_3^*] = 1.07 \times 10^{-5}$ M ; $[Sr^{2+}] = 7.15 \times 10^{-5}$ M ; $[Ca^{2-}] = 0$
Lessivage – Leaching (Durée – Duration $= 10t_0$)	$[H_2CO_3^*]$ 1.07×10^{-5} M ; $[Sr^{2+}]$ 0 ; $[Ca^{2+}]$ 4.63×10^{-3} M

lieu lors de cette expérience [9] entraînent un pic de Sr^{2+} retardé et amplifié par rapport à un traceur injecté dans les mêmes conditions (Fig. 1). Un calcul de référence est effectué à l'aide d'un maillage et de pas de temps très fins. Les courbes d'élution de Sr^{2+} sont alors identiques, que le calcul soit fait par un schéma NI ou par un schéma I.

En augmentant la taille des mailles et le pas de temps, il n'est plus possible de simuler correctement le pic de Sr^{2+} (Fig. 1). Il s'en suit une baisse de la concentration maximale atteinte ainsi qu'un étalement du pic. En revanche, le calcul du transport du traceur par la combinaison EFDMH conduit à une courbe d'élution identique, pour les deux discrétisations. Ainsi, une bonne modélisation du transport d'un traceur ne présage en rien de résultats corrects pour un soluté réactif. Avec cette méthode de résolution du transport, l'étalement du pic de Sr^{2+} n'est donc pas exclusivement dû aux phénomènes de diffusion numérique inévitables, liés à l'augmentation de la taille des mailles et du pas de temps. Avec une combinaison EFDMH, l'étalement du pic de Sr^{2+} est plus important avec le schéma NI qu'avec le schéma I. Les itérations entre les opérateurs de dispersion et de chimie conduisent à une meilleure approximation des transferts réactifs. Pour comparaison, l'utilisation d'une méthode plus classique, un schéma I associé à une résolution du transport par différences finies implicites en temps, montre l'intérêt évident de mettre en œuvre une méthode de résolution de l'opérateur de transport efficace : la diffusion numérique mise en évidence sur la courbe d'élution du traceur (Fig. 1) se traduit par un étalement très important du pic de Sr^{2+}.

4. Conclusion

Nous avons montré qu'il est possible de simuler le transport de solutés réactifs en milieu poreux à l'aide d'un schéma itératif en résolvant l'opérateur de transport par une combinaison EFDMH. Cette approche permet d'assurer un couplage précis entre transport et chimie par le biais du schéma itératif et de bénéficier d'une méthode de résolution performante de l'opérateur de transport pour des écoulements allant des phénomènes convectifs purs à des phénomènes purement diffusifs.

Remerciements

Ce travail a été financé par le Programme Environnement, Vie et Société du CNRS. J.C. a bénéficié d'une bourse du Ministère de l'Education Nationale, de la Recherche et de la Technologie (1998–2001).

Références

[1] J. van der Lee, L. de Windt, Present state and future directions of modeling of geochemistry in hydrogeological systems, J. Contaminant Hydrol. 47 (2001) 265–282.
[2] G.T. Yeh, V.S. Tripathi, A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components, Water Resource Res. 25 (1989) 93–108.
[3] P. Siegel, R. Mosé, Ph. Ackerer, J. Jaffre, Solution of the advection dispersion equation using a combination of discontinuous and mixed finite elements, Internat. J. Numer. Methods Fluids 24 (1997) 595–613.
[4] A.J. Valocchi, Validity of the local equilibrium assumption for modeling sorbing solute transport through homogeneous soils, Water Resource Res. 21 (1985) 808–820.
[5] F.M.M. Morel, Principles of Aquatic Chemistry, Wiley, New York, 1983.
[6] A. Cederberg, R.L. Street, J.O. Leckie, A groundwater mass transport and equilibrium chemistry model for multicomponent systems, Water Resource Res. 21 (1985) 1095–1104.
[7] C.I. Steefel, K.T.B. McQuarrie, Approaches to modelling of reactive transport in porous media, in: P.C. Lichtner, C.I. Steefel, E.H. Oelkers (Eds.), Reactive Transport in Porous Media, in: Reviews in Mineralogy, Vol. 34, Mineralogical Society of America, Washington, 1996, pp. 82–129.
[8] L. Sigg, Ph. Behra, W. Stumm, Chimie des milieux aquatiques, 3e édition, Masson, Paris, 2000.
[9] F. Lefèvre, M. Sardin, D. Schweich, Migration of strontium in a clayey and calcareous sand soil: precipitation and ion exchange, J. Contaminant Hydrol. 13 (1993) 215–229.

Annexe 2. Operator-splitting procedures for reactive transport and comparison of mass balance errors

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Contaminant Hydrology 68 (2004) 239–268
JOURNAL OF Contaminant Hydrology
www.elsevier.com/locate/jconhyd

Operator-splitting procedures for reactive transport and comparison of mass balance errors

Jérôme Carrayrou[a], Robert Mosé[a,b], Philippe Behra[a,c,*]

[a] *Institut de Mécanique des Fluides et des Solides de l'Université Louis Pasteur, UMR 7507 Université Louis Pasteur-CNRS, 2 rue Boussingault, 67000 Strasbourg, France*
[b] *Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg, 1 quai Koch, 67000 Strasbourg, France*
[c] *Ecole Nationale Supérieure des Ingénieurs en Arts Chimiques et Technologiques, Laboratoire de Chimie Agro-Industrielle, UMR 1010 INRA/INP-ENSIACET, 118 route de Narbonne, 31077 Toulouse Cedex 4, France*

Received 25 June 2001; received in revised form 10 June 2003; accepted 20 June 2003

Abstract

Operator-splitting (OS) techniques are very attractive for numerical modelling of reactive transport, but they induce some errors. Considering reactive mass transport with reversible and irreversible reactions governed by a first-order rate law, we develop analytical solutions of the mass balance for the following operator-splitting schemes: standard sequential non-iterative (SNI), Strang-splitting SNI, standard sequential iterative (SI), extrapolating SI, and symmetric SI approaches. From these analytical solutions, the operator-splitting methods are compared with respect to mass balance errors and convergence rates independently of the techniques used for solving each operator. Dimensionless times, N_{OS}, are defined. They control mass balance errors and convergence rates. The following order in terms of decreasing efficiency is proposed: symmetric SI, Strang-splitting SNI, standard SNI, extrapolating SI and standard SI schemes. The symmetric SI scheme does not induce any operator-splitting errors, the Strang-splitting SNI appears to be $O(N_{OS}^2)$ accurate, and the other schemes are first-order accurate.

Keywords: Reactive transport; Operator-splitting error; Non-iterative approach; Iterative approach; First-order kinetics

* Corresponding author. Ecole Nationale Supérieure des Ingénieurs en Arts Chimiques et Technologiques, Laboratoire de Chimie Agro-Industrielle, UMR 1010 INRA/INP-ENSIACET, 118 route de Narbonne, 31077 Toulouse Cedex 4, France. Tel.: +33-562-885-667; fax: +33-562-885-730.
E-mail address: Philippe.Behra@ensiacet.fr (Ph. Behra).

0169-7722/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0169-7722(03)00141-4

1. Introduction

In the last decade, reactive transport has been widely considered as a major topic in many disciplines such as fluid mechanics, combustion, chemical engineering, aquatic chemistry or earth sciences. Due to the lack of analytical solutions and to the nonlinearity of chemical processes, multicomponent reactive transport requires significant computational effort. Operator-splitting (OS) techniques allow the use of different and specific methods for solving both the transport and the chemical set of equations. Moreover, these techniques provide obvious opportunities for parallel computation (Hundsdorfer and Verwer, 1995). OS methods offer two distinct approaches. In a sequential non-iterative approach (SNI), the computational procedure of resolution for one time step is a sequence of one transport step followed by one chemical step (Walter et al., 1994; Hundsdorfer and Verwer, 1995; Appelo et al., 1997). This classical SNI approach can be modified by using two time steps as it is done in the Strang-splitting method (Hundsdorfer and Verwer, 1995; Steefel and MacQuarrie, 1996; Strang, 1968; Zysset et al., 1994). On the other hand, in the sequential iterative approach (SI), iterations are performed between transport and chemistry up to convergence for each time step (Hundsdorfer and Verwer, 1995; Leeming et al., 1998).

However, although OS methods show many advantages, the splitting procedure generates errors, which have been studied by several authors. Valocchi and Malmstead (1992) have calculated them by using an analytical solution for 1D transport with first-order radioactive decay and a flux boundary condition. They have shown that a standard SNI scheme induces an $O(\Delta t)$ error. They have proposed an alternating SNI scheme, which leads to an $O(\Delta t^2)$ error. Kaluarachchi and Morshed (1995a,b) have shown that the results obtained by Valocchi and Malmstead (1992) are not restricted to the first-order radioactive decay and flux boundary condition. These authors have analysed the error for a first-order kinetic reaction with flux or concentration boundary conditions, and for the transport of two species coupled with a Monod rate law. They have extended the results obtained by Valocchi and Malmstead (1992) and concluded that an $O(\Delta t)$ error occurs for the standard SNI scheme and an $O(\Delta t^2)$ error for the alternating SNI scheme. Barry et al. (1996) have analysed the splitting error for linear sorption, radioactive decay and interphase mass transfer chemistry in an arbitrary spatial domain. Additional numerical diffusion is introduced for a standard SNI scheme, if the initial concentrations in the chemical step are those coming from the beginning of the time step. On the other hand, the use of the calculated concentrations after the transport step increases the mass balance errors. They have also shown that the standard SNI leads to an $O(\Delta t)$ or an $O(\Delta t^2)$ error depending on the kind of chemistry and boundary conditions which are taken into account. Moreover, their results have suggested that the alternating SNI scheme does not lead to an $O(\Delta t^2)$ error in all cases. For a nonlinear equilibrium system with a single species system and for two competing species system for the standard SNI approach, Barry et al. (1997) have concluded that an $O(\Delta t)$ splitting error occurs for these conditions. For the standard (Hundsdorfer and Verwer, 1995) and the Strang-splitting (Lanser and Verwer, 1998) SNI schemes, the global errors in OS procedures are a combination of time-dependent and space-dependent errors. McRae et al. (1982) state that "there are two sources of error. . . . —the intrinsic error involved with OS and the discretisation errors associated with the operator approximations".

In this paper, our objective is to compare the intrinsic operator-splitting errors for the SNI and the SI strategies for the solution of a reactive transport equation with a kinetic formulation of chemical phenomenon. By using a mass balance formulation and a uniform monodirectional transport with fixed flux boundary conditions, each operator can be exactly solved under any initial conditions: reactive transport, conservative transport and batch chemistry operators. The discretisation errors associated with the operator approximation are therefore not discussed in this study. The mass balance formulation thus leads to a resolution without space discretisation, allowing the study of the intrinsic operator-splitting error alone. Analytical expressions of mass balance errors for non-iterative and iterative schemes are developed. For each studied scheme, the solution and mass balance errors are given at every time step. Moreover, we provide the convergence rate of the iterative schemes. Within the limit of the studied reactive transport problem, precision and efficiency of the different schemes will be studied in order to help select the most efficient scheme.

2. Review of the operator-splitting schemes

The transport of Nc reactive species is described by the set of Nc reactive transport equations (Eq. (1)), with the given boundary and initial conditions:

$$\frac{\partial c_i}{\partial t} = L(c_i, x, t) + f_i(c_1, \ldots, c_i, \ldots, c_{\mathrm{Nc}}) \; i = 1, \ldots, \mathrm{Nc} \tag{1}$$

where L is a transport operator, representing advection and dispersion, and f_i represents the chemistry influence over the i^{th} species, which depends on the local concentration of the other species.

In this study, a flux condition, $\Phi_i(t)$, is imposed on the upstream boundary of a semi-infinite domain:

$$\begin{cases} \mathrm{Flux}(t, x = 0) = \Phi_i(t) \\ \mathrm{Flux}(t, x = \infty) = 0 \end{cases} \tag{2}$$

The global approach for solving Eq. (1) over a discrete domain implies the computation of a very large system of equations the size of which is the number of species Nc × the number of nodes or cells Nn. A standard OS technique requires the computation of Nc independent transport systems of size Nn, because the splitting technique makes the transport of the i^{th} species independent of the other species, plus the computation of Nn independent chemistry systems of size Nc, because the splitting technique makes chemistry at one node independent of chemistry at the other nodes. Parallel computation of the problem is easier for OS. Moreover, the separation of the transport and the chemistry operators makes the use of highly specific methods for each operator possible. Several methods have been used for solving the transport equation (finite-difference or finite-element methods (Zysset et al., 1994; Šimůnek and Suarez, 1994), stochastic-advective

transport (Ginn, 2001), mixing cells (Jauzein et al., 1989; Leeming et al., 1998)) and the chemistry operator (predictor–corrector (Zysset et al., 1994), Gear's method (Hesstvedt et al., 1978), Rosenbrock solvers (Sandu et al., 1997)). Errors induced by the resolution of Eq. (1) through OS techniques are the combination of errors from both the transport and chemistry operators and the intrinsic OS errors. In order to eliminate the errors from both the transport and chemistry operators, we apply the global mass balance as proposed by Valocchi and Malmstead (1992). Moreover, this gives a generalisation of the results independent of the transport and chemistry operator computation. In the case of a linear relationship for describing chemistry, integration of the system (Eq. (1)) with respect to the system (Eq. (2)) over all the domain provides the following mass balance expression:

$$\frac{\mathrm{d}M_i}{\mathrm{d}t} = \Phi_i(t) + f(M_1, \ldots, M_i, \ldots, M_{\mathrm{Nc}}) \quad i = 1, \ldots, \mathrm{Nc} \tag{3}$$

$$M_i(t=0) = M_i^0 \quad i = 1, \ldots, \mathrm{Nc} \tag{4}$$

where M_i is the total mass of species i in the domain and M_i^0 the mass at $t=0$.

2.1. Standard sequential non-iterative scheme

The standard SNI scheme has been used for multicomponent reactive transport in groundwater (Walter et al., 1994; Appelo et al., 1997) and in the atmosphere (Hundsdorfer, 1996). From $n\Delta t$ to $(n+1)\Delta t$, this scheme is a combination of one nonreactive transport step followed by a batch chemistry step. The transport step requires the solution of the Nc following independent equations (Eq. (5)), which yields the intermediate solution M_i^* (Eq. (6)).

$$\frac{\mathrm{d}M_i^*}{\mathrm{d}t} = \Phi_i(t) \tag{5}$$

$M_{\mathrm{SS}i}^n$ is the total mass of species i provided by the standard SNI scheme at time $n\Delta t$ and the initial condition is: $M_i^*(n\Delta t) = M_{\mathrm{SS}i}^n$.

$$M_i^*[(n+1)\Delta t] = M_{\mathrm{SS}_i}^n + \int_{n\Delta t}^{(n+1)\Delta t} \Phi_i(t)\mathrm{d}t \tag{6}$$

The chemistry operator is then solved:

$$\frac{\mathrm{d}M_i^{**}}{\mathrm{d}t} = f_i(M_1, \ldots, M_i, \ldots, M_{\mathrm{Nc}}) \tag{7}$$

The initial condition at $n\Delta t$ is the intermediate solution: $M_i^{**}(n\Delta t) = M_i^*[(n+1)\Delta t]$. It yields the solution at time step $(n+1)$:

$$M_{\mathrm{SS}_i}^{n+1} = M_i^{**}[(n+1)\Delta t] = M_i^*[(n+1)\Delta t] + \int_{n\Delta t}^{(n+1)\Delta t} f_i(M_1, \ldots, M_i, \ldots, M_{\mathrm{Nc}})\mathrm{d}t \tag{8}$$

Using the standard SNI scheme for solving advection–dispersion reaction equations leads first to the solution of an advection–dispersion equation for each species i:

$$\frac{\boldsymbol{C}_i^*[(n+1)\Delta t] - C_{\mathrm{SS}_i}^n}{\Delta t} = -u\ \nabla \boldsymbol{C}_i + \nabla(\boldsymbol{D}\ \nabla \boldsymbol{C}_i) \tag{9}$$

where $\boldsymbol{C}_i$ is the vector of the concentration of species i at each node of the domain. The reaction equation is then solved for each species i at each node of the discretised domain:

$$\frac{\boldsymbol{C}_{\mathrm{SS}_i}^{n+1} - \boldsymbol{C}_i^*[(n+1)\Delta t]}{\Delta t} = f_i(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_i, \ldots, \boldsymbol{C}_{\mathrm{Nc}}) \tag{10}$$

The results of the analysis performed in this work are independent of the method used for solving each operator. The time discretisation scheme is therefore not specified in Eqs. (9) and (10).

2.2. Strang-splitting sequential non-iterative scheme

Proposed by Strang (1968), this scheme has been used by several authors (Zysset, 1993; Zysset et al., 1994; Hundsdorfer, 1996; Steefel and MacQuarrie, 1996). The Alternating SNI scheme studied by Kaluarachchi and Morshed (1995a,b) is equivalent to the Strang-splitting scheme with a half time step. Results obtained by these authors with the Alternating SNI scheme will be used here for the Strang-splitting scheme. $M_{\mathrm{ST}_i}^n$ is the total mass of species i given by the Strang-splitting SNI scheme at time $n\Delta t$.

In this procedure, the transport equation (Eq. (5)) is first solved over the half time step, i.e. from $n\Delta t$ to $(n+1/2)\Delta t$ to give the first intermediate solution, M_i^*. The initial condition at $n\Delta t$ is the solution of the previous time step: $M_i^*(n\Delta t) = M_{\mathrm{STi}}^n$:

$$M_i^*[(n+1/2)\Delta t] = M_{\mathrm{ST}_i}^n + \int_{n\Delta t}^{(n+1/2)\Delta t} \Phi_i(t)\mathrm{d}t \tag{11}$$

For $M_i^{**}(n\Delta t) = M_i^*[(n+1/2)\Delta t]$ as initial conditions, the chemistry operator (Eq. (7)) is solved over the entire time step to obtain the second intermediate solution M_i^{**} at $(n+1)\Delta t$:

$$M_i^{**}[(n+1)\Delta t] = M_i^*[(n+1/2)\Delta t] + \int_{n\Delta t}^{(n+1)\Delta t} f_i(M_1, \ldots, M_i \ldots, M_{\mathrm{Nc}})\mathrm{d}t \tag{12}$$

The solution of the Strang-splitting SNI scheme at $(n+1)\Delta t$ is then obtained after solving a second transport equation (Eq. (5)) from $(n+1/2)\Delta t$ to $(n+1)\Delta t$ with the second intermediate solution as initial condition, $M_i^{***}[(n+1/2)\Delta t] = M_i^{**}[(n+1)\Delta t]$:

$$M_{\mathrm{ST}_i}^{n+1} = M_i^{***}[(n+1)\Delta t] = M_i^{**}[(n+1)\Delta t] + \int_{(n+1/2)\Delta t}^{(n+1)\Delta t} \Phi_i(t)\mathrm{d}t \tag{13}$$

For each advection–dispersion reaction equation, the resolution of the advection–dispersion equation over the first half-time step is:

$$\frac{\boldsymbol{C}_i^*[(n+1/2)\Delta t] - C_{\mathrm{ST}_i}^n}{\Delta t/2} = -u\ \nabla \boldsymbol{C}_i + \nabla(\boldsymbol{D}\ \nabla \boldsymbol{C}_i) \tag{14}$$

The chemistry operator is solved at each node of the discretised domain over the entire time step for each species i:

$$\frac{\boldsymbol{C}_i^{**}[(n+1)\Delta t] - \boldsymbol{C}_i^*[(n+1/2)\Delta t]}{\Delta t} = f_i(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_i \ldots, \boldsymbol{C}_{\mathrm{Nc}}) \tag{15}$$

The second advection–dispersion equation is solved over the second half-time step:

$$\boldsymbol{C}_{\mathrm{ST}_i}^{n+1} - \frac{\boldsymbol{C}_i^{**}[(n+1)\Delta t]}{\Delta t/2} = -u\ \nabla \boldsymbol{C}_i + \nabla(\boldsymbol{D}\ \nabla \boldsymbol{C}_i) \tag{16}$$

2.3. Standard sequential iterative scheme

Several authors have used the Standard SI scheme, or some schemes very close to this one (Yeh and Tripathi, 1989; Hundsdorfer and Verwer, 1995; Leeming et al., 1998). $M_{\mathrm{IT}_i}^n$ is the total mass of species i given by the Standard SI scheme at time $n\Delta t$. For one time step, this scheme is decomposed in a conservative transport step with a chemical source–sink term. $M_{\mathrm{T}_i}^m$ is the mass used by the transport operator and $R_{\mathrm{C}_i}^m$ the chemistry source–sink term for the species i at iteration m:

$$\frac{\mathrm{d}M_{\mathrm{T}_i}^m}{\mathrm{d}t} = \Phi_i(t) - R_{\mathrm{C}_i}^{m-1} \tag{17}$$

$R_{\mathrm{C}_i}^m$ represents the chemistry operator $f_i(M_1 \ldots, M_i \ldots, M_{\mathrm{Nc}})$, but $R_{\mathrm{C}_i}^m$ is a constant (during one iteration) in the transport operator, whereas $f_i(M_1, \ldots, M_i, \ldots, M_{\mathrm{Nc}})$ is a system of ordinary differential equations. With the solution of the previous time step as initial condition, $M_{\mathrm{T}_i}^m(n\Delta t) = M_{\mathrm{IT}_i}^n$, the solution of the transport operator at iteration m is:

$$M_{\mathrm{T}_i}^m[(n+1)\Delta t] = M_{\mathrm{IT}_i}^n + \int_{n\Delta t}^{(n+1)\Delta t} \Phi_i(t)\mathrm{d}t - R_{\mathrm{C}_i}^{m-1}\Delta t \tag{18}$$

The batch chemical operator is solved over the time step with $M_{\mathrm{C}_i}^m$ the mass of species i calculated by the chemistry operator at iteration m:

$$\frac{\mathrm{d}M_{\mathrm{C}_i}^m}{\mathrm{d}t} = f_i(M_1, \ldots, M_i, \ldots, M_{\mathrm{Nc}}) \tag{19}$$

To define the initial condition of the chemistry operator, $M^m_{C_i}(n\Delta t)=M^m_{T_i}[(n+1)\Delta t]$, we obtain the solution:

$$M^m_{C_i}[(n+1)\Delta t] = M^m_{T_i}[(n+1)\Delta t] + \int_{n\Delta t}^{(n+1)\Delta t} f_i(M_1,\ldots,M_i,\ldots,M_{Nc})\mathrm{d}t \tag{20}$$

which allows the updating of the chemistry source–sink term for the next iteration:

$$R^m_{C_i} = \frac{M^m_{T_i}[(n+1)\Delta t] - M^m_{C_i}[(n+1)\Delta t]}{\Delta t} \tag{21}$$

Both steps are repeated until the convergence limit ε is reached:

$$\mathrm{CV}^m_{\mathrm{IS}_i} = \frac{\left|M^{n+1,m}_{T_i} - M^{n+1,m-1}_{T_i}\right|}{M^{n+1,m}_{T_i}} \le \varepsilon \tag{22}$$

If the algorithm converges, its limit is the solution of the standard SI at $t=(n+1)\Delta t$:

$$M^{n+1}_{\mathrm{IT}_i} = \lim_{m\to\infty} M^m_{T_i}[(n+1)\Delta t] \tag{23}$$

When the standard SI scheme is used, the advection–dispersion-reaction equation is solved using the following algorithm. The chemistry source–sink term is added to the advection–dispersion equation:

$$\frac{\boldsymbol{C}^m_{T_i}[(n+1)\Delta t] - C^n_{\mathrm{SS}_i}}{\Delta t} = -u\ \nabla \boldsymbol{C}_{T_i} + \nabla(\boldsymbol{D}\ \nabla \boldsymbol{C}_{T_i}) - \boldsymbol{R}^{m-1}_{C_i} \tag{24}$$

where $\boldsymbol{R}^{m-1}_{C_i}$ is the vector of the chemistry source–sink terms at each node of the domain. The solution of the chemistry operator (Eq. (25)) provides the value of the chemistry source–sink term at the next iteration (Eq. (26)):

$$\frac{\boldsymbol{C}^m_{C_i}[(n+1)\Delta t] - \boldsymbol{C}^m_{T_i}[(n+1)\Delta t]}{\Delta t} = f_i(\boldsymbol{C}_{C_1},\ldots,\boldsymbol{C}_{C_i},\ldots,\boldsymbol{C}_{CN_c}) \tag{25}$$

$$\boldsymbol{R}^m_{C_i} = \frac{\boldsymbol{C}^m_{T_i}[(n+1)\Delta t] - \boldsymbol{C}^m_{C_i}[(n+1)\Delta t]}{\Delta t} \tag{26}$$

The convergence criterion must be obeyed at every node of the domain and for all species i:

$$\mathrm{CV}^m_{\mathrm{IS}_i} = \max \frac{\left|\boldsymbol{C}^{n+1,m}_{T_i} - \boldsymbol{C}^{n+1,m-1}_{T_i}\right|}{\boldsymbol{C}^{n+1,m}_{T_i}} \le \varepsilon \tag{27}$$

Special treatment of the convergence criterion is required for nodes where C_{Ti}^{n+1} is zero or negligible.

2.4. Extrapolating sequential iterative scheme

Some authors such as Cederberg et al. (1985) have proposed the extrapolating SI scheme which is a modification of the standard SI scheme. $M_{IE_i}^n$ is the total mass of species i given by the extrapolating SI scheme at time $n\Delta t$. A standard SI scheme is performed for the first half of the time step until convergence is reached giving the intermediate solution $M_{IT_i}^*[(n+1/2)\Delta t]$. Then the solution at the end of the time step is linearly extrapolated through the intermediate solution:

$$M_{IE_i}^{n+1} = 2\ M_{IT_i}^*[(n+1/2)\Delta t] - M_{IE_i}^n \tag{28}$$

The standard SI scheme, Eqs. (24)–(27), is used with a half time step for the solution of the advection–dispersion reaction equation and yields $C_{IT_i}^*[(n+1/2)\Delta t]$. The extrapolation procedure provides the concentration at the end of the time step at each node of the domain:

$$C_{IE_i}^{n+1} = 2\ C_{IT_i}^*[(n+1/2)\Delta t] - C_{IE_i}^n \tag{29}$$

2.5. Symmetric sequential iterative scheme

Herzer (1989) has described an iterative scheme with two source–sink terms. A chemistry source–sink term is introduced into the transport operator as for the other iterative schemes and a transport source–sink term is introduced into the chemistry operator (noted R_T^m). Both operators are then written in a symmetric manner. Many other authors have used a scheme based on the same concept (Zysset, 1993; Hundsdorfer and Verwer, 1995; Kanney et al., 2003). The algorithm includes two steps for one iteration. $M_{IR_i}^n$ is the total mass of species i given by the symmetric SI scheme at time $n\Delta t$. The transport operator (Eq. (17)) is solved with the solution of the previous time step as initial condition, $M_{T_i}^m(n\Delta t) = M_{IR_i}^n$, yielding the transport solution (Eq. (18)). Then the transport source–sink term is calculated:

$$R_{T_i}^m = \frac{M_{IR_i}^n - M_{T_i}^m[(n+1)\Delta t]}{\Delta t} - R_{C_i}^{m-1} \tag{30}$$

The chemistry operator is written with a constant transport source–sink term:

$$\frac{dM_{C_i}^m}{dt} = f_i(M_1, \ldots, M_i, \ldots, M_{Nc}) - R_{T_i}^m \tag{31}$$

With the solution of the previous time step as initial condition, $M_{C_i}^m(n\Delta t) = M_{IR_i}^n$, the chemistry solution is:

$$M_{C_i}^m[(n+1)\Delta t] = M_{IR_i}^n + \int_{n\Delta t}^{(n+1)\Delta t} f_i(M_1, \ldots, M_i, \ldots, M_{Nc})dt - R_{T_i}^m \Delta t \tag{32}$$

The chemistry source–sink term is calculated again:

$$R_{C_i}^m = \frac{M_{IR_i}^n - M_{C_i}^m[(n+1)\Delta t]}{\Delta t} - R_{T_i}^m \tag{33}$$

Iterations between transport and chemistry operators are performed until convergence is reached:

$$CV_{IR_i}^m = \frac{\left|M_{T_i}^m[(n+1)\Delta t] - M_{C_i}^m[(n+1)\Delta t]\right|}{M_{T_i}^m[(n+1)\Delta t] + M_{C_i}^m[(n+1)\Delta t]} \leq \varepsilon \tag{34}$$

Both transport and chemistry operators provide the solution of the symmetric SI scheme. In order to preserve the symmetry of this scheme, solutions of both operators are used in the rationale for the mathematical form of the convergence criterion (Eq. (34)). If the algorithm converges, its limit is the solution of the symmetric SI scheme at $t=(n+1)\Delta t$:

$$M_{IR_i}^{n+1} = \lim_{m\to\infty} M_{T_i}^m[(n+1)\Delta t] = \lim_{m\to\infty} M_{C_i}^m[(n+1)\Delta t] \tag{35}$$

For the solution of the advection–dispersion-reaction equation, the advection–dispersion equation with the chemistry source–sink term (Eq. (24)) is solved for each species i. The vector of the transport source–sink term $\boldsymbol{R}_{T_i}^m$ at each node of the domain is calculated for iteration m:

$$\boldsymbol{R}_{T_i}^m = \frac{\boldsymbol{C}_{IR_i}^n - \boldsymbol{C}_{T_i}^m[(n+1)\Delta t]}{\Delta t} - \boldsymbol{R}_{C_i}^{m-1} \tag{36}$$

The transport source–sink term vector is added to the chemistry operator, which is solved at each node of the domain for all species:

$$\frac{\boldsymbol{C}_{C_i}^m[(n+1)\Delta t] - \boldsymbol{C}_{IR_i}^n}{\Delta t} = f_i(\boldsymbol{C}_{C_1}, \ldots, \boldsymbol{C}_{C_i}, \ldots, \boldsymbol{C}_{CN_c}) - \boldsymbol{R}_{T_i}^m \tag{37}$$

The solution of the chemistry operator is then used for the computation of the chemistry source–sink term vector at iteration m:

$$\boldsymbol{R}_{C_i}^m = \frac{\boldsymbol{C}_{T_i}^m[(n+1)\Delta t] - \boldsymbol{C}_{IR_i}^n}{\Delta t} - \boldsymbol{R}_{T_i}^m \tag{38}$$

The convergence criterion is defined as:

$$CV_{IR_i}^m = \max \frac{\left|\boldsymbol{C}_{T_i}^m[(n+1)\Delta t] - C_{C_i}^m[(n+1)\Delta t]\right|}{\boldsymbol{C}_{T_i}^m[(n+1)\ \Delta t] + \boldsymbol{C}_{C_i}^m[(n+1)\Delta t]} \leq \varepsilon \tag{39}$$

This criterion is imposed at each node of the domain and for all the species i, with a special consideration to nodes where solution of the transport and chemistry operators are zero or negligible.

3. Calculations of mass balance errors

3.1. Analytical mass balance for first-order irreversible and reversible reaction

The OS mass balance errors has been analysed with a first-order decay reaction for one species (Valocchi and Malmstead, 1992), but only non-iterative schemes have been studied. In this section, irreversible (one species) and reversible (two species) first-order rate laws are used to calculate analytical mass balance solution for both non-iterative and iterative OS procedures.

The reversible first-order reaction is written as:

$$c_1 \underset{k_1}{\overset{k_2}{\leftrightarrows}} c_2 \tag{40}$$

where k_1 and k_2 are the rate constants.

For simplification, the flux at the inlet of the domain is assumed to be constant over time. The mass balance evolution can be written with $i=1$ or 2 and $j=2$ or 1, with $j \neq i$:

$$\begin{cases} \dfrac{\mathrm{d}M_i}{\mathrm{d}t} = \Phi_i - k_i \;\; M_i + k_j \;\; M_j \\ \\ M_i(t=0) = M_i^0 \end{cases} \tag{41}$$

3.2. Exact solution

By adding Eq. (41) for $i=1$, and Eq. (41) for $i=2$, a differential equation for the total mass of species 1 and 2 is obtained:

$$M_{\mathrm{tot}}(t) = M_{\mathrm{tot}}^0 + \Phi_{\mathrm{tot}} \;\; t \tag{42}$$

with $M_{\mathrm{tot}}^0 = M_i^0 + M_j^0$ and $\Phi_{\mathrm{tot}} = \Phi_i + \Phi_j$. From Eqs. (41) and (42), a new relationship is obtained that depends on the species i, with $k_{\mathrm{tot}} = k_i + k_j$:

$$\begin{cases} \dfrac{\mathrm{d}M_i}{\mathrm{d}t} = \Phi_i + k_j \;\; M_{\mathrm{tot}}^0 + k_j \;\; \Phi_{\mathrm{tot}} \;\; t - k_{\mathrm{tot}} \;\; M_i \\ \\ M_i(t=0) = M_i^0 \end{cases} \tag{43}$$

M_{EX}, the exact solution of Eq. (41) is:

$$M_{\mathrm{EX}_i}(t) = M_i^0 + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} \right) \;\; k_{\mathrm{tot}} t + \left[\frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} - \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} \right] (1 - e^{-k_{\mathrm{tot}} t}) \tag{44}$$

We note for a variable X:

$$\Delta_i(kX) = k_j X_j - k_i X_i \tag{45}$$

$\Delta_i(kX)$ represents the deviation from thermodynamic equilibrium of the variable X for the formation of species i. Indeed, at equilibrium, $\Delta_i(kX)$ is zero. If X_i is above (respectively below), its equilibrium value, $\Delta_i(kX)$, is negative (respectively positive).

For one species only and $M_i^0 = 0$, the exact solution of a first-order irreversible reaction is:

$$M_{\mathrm{EX}}(t) = \frac{\Phi}{k}\left(1 - e^{-kt}\right) \tag{46}$$

3.3. Standard SNI

The Standard SNI algorithm provided by Eqs. (5)–(8) is used for solving the reactive transport problem (Eq. (41)). The solution of the chemistry operator is performed using the same method as for the exact global solution, described by Eqs. (41)–(44). This yields the recursive form (Eq. (47)):

$$M_{\mathrm{SS}_i}^{n+1} = M_{\mathrm{SS}_i}^{n} + \frac{\Phi_i}{k_{\mathrm{tot}}} k_{\mathrm{tot}}\Delta t + \left(\frac{\Delta_i(kM_{\mathrm{SS}}^n)}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} k_{\mathrm{tot}}\Delta t\right)\left(1 - e^{-k_{\mathrm{tot}}\Delta t}\right) \tag{47}$$

The solution given by the Standard SNI scheme is written in the explicit form (Eq. (48)) (see details of the solution in Appendix A):

$$M_{\mathrm{SS}_i}^{n} = M_i^0 + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right) n k_{\mathrm{tot}}\Delta t + \left(\frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} - \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\frac{e^{-k_{\mathrm{tot}}\Delta t}}{1 - e^{-k_{\mathrm{tot}}\Delta t}} k_{\mathrm{tot}}\Delta t\right) \times \left(1 - e^{-nk_{\mathrm{tot}}\Delta t}\right) \tag{48}$$

For a standard SNI scheme, Valocchi and Malmstead (1992) have proposed the solution for one species and zero initial mass. They have reported the mass and the relative mass balance error for a standard SNI scheme, M_{SS}, after one time step. Kaluarachchi and Morshed (1995a) have calculated the mass for a standard SNI scheme after n time steps (Eq. (49)), and have given the analytical value of the mass balance error as:

$$M_{\mathrm{SS}}^{n} = \frac{\Phi}{k} k\Delta t \left(\frac{e^{-k\Delta t}}{1 - e^{-k\Delta t}}\right)\left(1 - e^{-nk\Delta t}\right) \tag{49}$$

3.4. Strang-splitting SNI

The Strang-splitting algorithm, detailed by Eqs. (11)–(13), yields the recursive solution:

$$M_{\mathrm{ST}_i}^{n+1} = M_{\mathrm{ST}_i}^{n} + \frac{\Phi_i}{k_{\mathrm{tot}}} k_{\mathrm{tot}}\Delta t + \left(\frac{\Delta_i(kM_{\mathrm{ST}}^n)}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\frac{k_{\mathrm{tot}}\Delta t}{2}\right)\left(1 - e^{-k_{\mathrm{tot}}\Delta t}\right) \tag{50}$$

The explicit form is (see Appendix B):

$$M_{\mathrm{ST}_i}^n = M_i^0 + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right) n k_{\mathrm{tot}} \Delta t + \left(\frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} - \frac{1}{2}\frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\frac{1+e^{-k_{\mathrm{tot}}\Delta t}}{1-e^{-k_{\mathrm{tot}}\Delta t}} k_{\mathrm{tot}}\Delta t\right) \times \left(1 - e^{-n k_{\mathrm{tot}}\Delta t}\right) \tag{51}$$

Kaluarachchi and Morshed (1995a) have calculated the mass of the Alternating SNI scheme and the mass balance error after even or odd time steps. The Alternating SNI scheme is identical to the Strang-splitting SNI with a doubled time step. The total mass balance solution (Eq. (52)) obtained with a Strang-splitting SNI scheme after n time steps is:

$$M_{\mathrm{ST}}^n = \frac{\Phi}{k}\left(1 + e^{-k\Delta t}\right)\frac{1 - e^{-nk\Delta t}}{1 - e^{-k\Delta t}} k \frac{\Delta t}{2} \tag{52}$$

3.5. Standard SI

The standard SI algorithm described by Eqs. (17)–(19) is used for solving the reactive transport problem (Eq. (41)). This yields:

$$M_{\mathrm{C}_i}^m[(n+1)\Delta t] = M_{\mathrm{T}_i}^m[(n+1)\Delta t] + \frac{\Delta_i(kM_{\mathrm{T}}^m[(n+1)\Delta t])}{k_{\mathrm{tot}}}\left(1 - e^{-k_{\mathrm{tot}}\Delta t}\right) \tag{53}$$

which is the recursive solution of the chemistry operator (Eq. (20)). The chemistry source–sink term defined in Eq. (21) can then be written as:

$$R_{\mathrm{C}_i}^{m+1} = -\frac{\Delta_i(kM_{\mathrm{T}}^m[(n+1)\Delta t])}{k_{\mathrm{tot}}} \frac{1 - e^{-k_{\mathrm{tot}}\Delta t}}{\Delta t} \tag{54}$$

This yields the recursive form of the transport operator (Eq. (18)):

$$M_{\mathrm{T}_i}^{m+1}[(n+1)\Delta t] = M_{\mathrm{IS}_i}^n + \frac{\Phi_i}{k_{\mathrm{tot}}} k_{\mathrm{tot}}\Delta t + \frac{\Delta_i(kM_{\mathrm{T}}^m[(n+1)\Delta t])}{k_{\mathrm{tot}}}\left(1 - e^{-k_{\mathrm{tot}}\Delta t}\right) \tag{55}$$

Writing Eq. (55) for i and j yields a recursive formulation of the $\Delta_i(kM_{\mathrm{T}}^m[(n+1)\Delta t]$ series (see Appendix C). The explicit form of this series is used to obtain:

$$M_{\mathrm{T}_i}^{m+1}[(n+1)\Delta t] = M_{\mathrm{IS}_i}^n + \frac{\Phi_i}{k_{\mathrm{tot}}} k_{\mathrm{tot}}\Delta t + \frac{\Delta_i(kM_{\mathrm{IS}}^n) + \Delta_i(k\Phi)\Delta t}{k_{\mathrm{tot}}} \times \frac{1 - e^{-k_{\mathrm{tot}}\Delta t}}{2 - e^{-k_{\mathrm{tot}}\Delta t}}\left[1 - (-1)^m\left(1 - e^{-k_{\mathrm{tot}}\Delta t}\right)^m\right] \tag{56}$$

The series of the mass balance solution given at each iteration by the standard SI scheme converges if $m \longrightarrow \infty$. The solution for $m \longrightarrow \infty$ is used to calculate the solution of the standard SI at the next time step:

$$M_{\mathrm{IS}_i}^{n+1} = M_{\mathrm{IS}_i}^{n} + \frac{\Phi_i}{k_{\mathrm{tot}}} k_{\mathrm{tot}} \Delta t + \left(\frac{\Delta_i(kM_{\mathrm{IS}}^{n})}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} k_{\mathrm{T}} \Delta t \right) \frac{1 - e^{-k_{\mathrm{tot}}\Delta t}}{2 - e^{-k_{\mathrm{tot}}\Delta t}} \tag{57}$$

This solution can be expressed in an explicit form (see Appendix C):

$$M_{\mathrm{IS}_i}^{n} = M_i^0 + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} \right) n k_{\mathrm{tot}} \Delta t + \left(\frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} - \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} \frac{k_{\mathrm{tot}}\Delta t}{1 - e^{-k_{\mathrm{tot}}\Delta t}} \right) \times \left[1 - \left(\frac{1}{2 - e^{-k_{\mathrm{tot}}\Delta t}} \right)^n \right] \tag{58}$$

For one species only, Eq. (58) simplifies to the form:

$$M_{\mathrm{IS}}^{n} = \frac{\Phi}{k} \left(\frac{1}{1 - e^{-k\Delta t}} \right) \left[1 - \left(\frac{1}{2 - e^{-k\Delta t}} \right)^n \right] k\Delta t \tag{59}$$

From these developments, we show that the Standard SI scheme does not yield exact results. Convergence of the iterative procedure does not imply that the mass balance is exact.

3.6. Extrapolating SI

The first part of the extrapolating SI algorithm is a standard SI scheme applied over the half time step until convergence. The iteration procedure and the convergence criterion used for the extrapolating SI scheme are the same as those used for the standard SI, Eqs. (55) and (57), but with a half time step. This gives the intermediate solution:

$$M_{\mathrm{IE}_i}^{*}[(n + 1/2)\Delta t] = M_{\mathrm{IE}_i}^{n} + \frac{\Phi_i}{k_{\mathrm{tot}}} k_{\mathrm{tot}} \frac{\Delta t}{2} + \left(\frac{\Delta_i(kM_{\mathrm{IE}}^{n})}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} k_{\mathrm{tot}} \frac{\Delta t}{2} \right) \times \frac{1 - e^{-k_{\mathrm{tot}}\frac{\Delta t}{2}}}{2 - e^{-k_{\mathrm{tot}}\frac{\Delta t}{2}}} \tag{60}$$

Eq. (28) is formulated more explicit by using the value of the intermediate solution (Eq. (60)). This yields the recursive form of the solution of the extrapolating SI at the next time step:

$$M_{\mathrm{IE}_i}^{n+1} = M_{\mathrm{IE}_i}^{n} + \frac{\Phi_i}{k_{\mathrm{tot}}} k_{\mathrm{tot}} \Delta t + 2 \left(\frac{\Delta_i(kM_{\mathrm{IE}}^{n})}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} k_{\mathrm{tot}} \frac{\Delta t}{2} \right) \frac{1 - e^{-k_{\mathrm{tot}}\frac{\Delta t}{2}}}{2 - e^{-k_{\mathrm{tot}}\frac{\Delta t}{2}}} \tag{61}$$

Calculations to obtain the explicit form (Eq. (62)) from Eq. (61) are very similar to these used for the standard SI scheme (see details in Appendix C):

$$M_{\mathrm{IE}_i}^n = M_i^0 + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right) n k_{\mathrm{tot}} \Delta t + \left(\frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} - \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2} \frac{k_{\mathrm{tot}} \frac{\Delta t}{2}}{1 - e^{-k_{\mathrm{tot}} \frac{\Delta t}{2}}}\right) \times \left[1 - \left(\frac{e^{-k_{\mathrm{tot}} \frac{\Delta t}{2}}}{2 - e^{-k_{\mathrm{tot}} \frac{\Delta t}{2}}}\right)^n\right] \tag{62}$$

For the transport of a single species, Eq. (61) can be simplified. In this case, the mass balance solution (Eq. (63)) obtained by the extrapolating SI scheme at each time step is:

$$M_{\mathrm{IE}}^n = \frac{\Phi}{k} \frac{1}{1 - e^{-k\frac{\Delta t}{2}}} \left[1 - \left(\frac{e^{-k\frac{\Delta t}{2}}}{2 - e^{-k\frac{\Delta t}{2}}}\right)^n\right] k \frac{\Delta t}{2} \tag{63}$$

3.7. Symmetric SI

The solution of transport step (Eq. (18)) of the symmetric SI scheme is:

$$M_{\mathrm{T}_i}^m[(n+1)\Delta t] = \Phi_i \Delta t - R_{C_i}^{m-1} \Delta t \tag{64}$$

From the definition of the transport source–sink term (Eq. (30)) applied to a constant flux boundary condition, the explicit formulation for the first iteration is:

$$R_{\mathrm{T}_i}^1 = -\Phi_i \tag{65}$$

If this explicit form is introduced into the chemistry equation (Eq. (31)) used for the symmetric SI algorithm, the same equation than for the exact reactive transport equation (Eq. (41)) is obtained:

$$\begin{cases} \dfrac{\mathrm{d}M_i}{\mathrm{d}t} = -k_i \cdot M_i + k_j \cdot M_j + \Phi_i \\ M_i(t=0) = M_i^0 \end{cases} \tag{66}$$

If the solution of the symmetric SI scheme at $n\Delta t$ is the exact solution (this is true for $n=0$), the solution of the chemistry operator in the symmetric SI algorithm corresponds to the exact solution given by Eq. (44). From Eq. (33), the explicit formulation of the chemistry source–sink term can be derived as:

$$R_{C_i}^{m-1} = \frac{M_{\mathrm{IR}_i}^n - M_{\mathrm{EX}_i}^{n+1}}{\Delta t} - \Phi_i \tag{67}$$

Solving Eq. (17) with a chemistry source–sink term, we obtain the expression of the transport solution at the second iteration:

$$M_{\mathrm{IR}_i}^{m-2}[(n+1)\Delta t] = M_{\mathrm{EX}_i}^{n+1} \tag{68}$$

Then the transport and the chemistry operators yield the same solution, which is exact at the second iteration:

$$M_{\mathrm{IR}_i}^{n} = M_i^0 + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right) nk_{\mathrm{tot}}t + \left[\frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} - \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right]\left(1 - e^{-nk_{\mathrm{tot}}t}\right) \tag{69}$$

In the case of a constant flux boundary condition, the symmetric SI algorithm gives the exact mass balance at the second iteration, whatever the chemistry. Indeed, the formulation of the chemistry source–sink term (Eq. (67)) does not depend on the explicit form of the chemistry function $f_i(M_1,\ldots,M_i,\ldots,M_{\mathrm{Nc}})$. If only one species is taken into account, the mass balance is:

$$M_{\mathrm{IR}}^{n} = \frac{\Phi}{k}\left(1 - e^{-nk\Delta t}\right) \tag{70}$$

which implies a zero mass balance error.

3.7.1. Relative mass balance errors

Valocchi and Malmstead (1992) have studied the mass balance error in the case of one species with a first-order rate law. They have defined the mass balance error as the relative difference between the exact and calculated masses. Since the masses tend towards maximal values at steady state, those mass balance errors do not tend towards zero when time goes to infinity. For a reversible reaction, the masses increase continuously, nullifying the relative difference between the exact and calculated masses. The presence of initial mass also influences those mass balance errors. In order to eliminate the influence of the initial mass and the increase in the masses, the relative mass balance error in the case of two species is defined as:

$$E_i^n = 1 - \frac{M_i^n - \left[M_i^0 + \frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right) nk_{\mathrm{tot}}\Delta t\right]}{M_{\mathrm{EX}_i}^n - \left[M_i^0 + \frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right) nk_{\mathrm{tot}}\Delta t\right]} \tag{71}$$

where $M_i^0 + \frac{\Delta_i(kM^0)}{k_{\mathrm{tot}}} + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{tot}^2}\right) nk_{\mathrm{tot}}\Delta t$ is the amount of the species i in the domain at time $n\Delta t$, assuming that the thermodynamic equilibrium is instantaneously obtained. The deviations from the thermodynamic equilibrium are also compared between the exact solution and the solution calculated by operator-splitting methods. For one species only, the relative mass balance error (Eq. (71)) is the one defined by Valocchi and Malmstead (1992):

$$E^n = 1 - \frac{M^n}{M_{\mathrm{EX}}^n} \tag{72}$$

Definition of the relative mass balance error (Eq. (71)) is consistent with the classical definition (Eq. (72)), since for one species, at equilibrium $M=0$. For a first-order rate, explicit formulations of the mass balance solutions can be calculated for each studied scheme. The relative mass balance error definition (Eq. (71)) is thus used with one or two species (see Tables 1 and 2).

3.7.2. Dimensionless times

From now on, the two following dimensionless times are used (see Tables 1 and 2), either:

$$N_{OS1} = k\Delta t \tag{73}$$

for one species and an irreversible reaction with a first-order rate law, or:

$$N_{OS2} = k_{tot}\Delta t \tag{74}$$

for two species and a reversible reaction with a first-order rate law.

N_{OS1} and N_{OS2} control the mass calculated by OS procedures as shown in mass balance (Eqs. (48) and (49)) for the standard SNI, Eqs. (51) and (52) for the Strang-splitting SNI, Eqs. (58) and (59) for the standard SI, Eqs. (62) and (63) for the extrapolating SI, and Eqs. (69) and (70) for the symmetric SI schemes. These dimensionless times provide

Table 1
Relative mass balance errors at steady-state condition for first-order rate law: (i) irreversible reaction, $i=1$ and $N_{OS1}=k\Delta t$; (ii) reversible reaction, $i=2$ and $N_{OS2}=k_{tot}\Delta t$

Scheme	Relative mass balance error	Taylor expansion
Standard SNI	$E_{SS} = 1 - \frac{e^{-N_{OSi}}}{1-e^{-N_{OSi}}} N_{OSi}$	$E_{SS} = \frac{1}{2} N_{OSi} + O(N_{OSi}^2)$
Strang-splitting SNI	$E_{ST} = 1 - \frac{1+e^{-N_{OSi}}}{1-e^{-N_{OSi}}} \frac{N_{OSi}}{2}$	$E_{ST} = -\frac{1}{12} N_{OSi}^2 + O(N_{OSi}^4)$
Standard SI	$E_{IS} = 1 - \frac{N_{OSi}}{1-e^{-N_{OSi}}}$	$E_{IS} = -\frac{1}{2} N_{OSi} + O(N_{OSi}^2)$
Extrapolating SI	$E_{IE} = 1 - \frac{1}{1-e^{\frac{N_{OSi}}{2}}} \frac{N_{OSi}}{2}$	$E_{IE} = -\frac{1}{4} N_{OSi} + O(N_{OSi}^2)$
Symmetric SI	$E_{IR} = 0$	$E_{IR} = 0$

Table 2
Relative mass balance errors evolution of first-order reaction for one or two species

Scheme	First-order reaction, one species	First-order reaction, two species
Standard SNI	$E_{SS} = 1 - \frac{e^{-N_{OS1}}}{1-e^{-N_{OS1}}} N_{OS1}$	$E_{SS}^n = 1 - \frac{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} e^{-nN_{OS1}} + \frac{\varepsilon}{1-e^{-N_{OS2}}} N_{OS2}(1-e^{-nN_{OS1}})}{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} e^{-nN_{OS1}} + (1-e^{-nN_{OS2}})}$
Strang-splitting SNI	$E_{ST} = 1 - \frac{1+e^{-N_{OS}}}{1-e^{-N_{OS}}} \frac{N_{OS1}}{2}$	$E_{ST}^n = 1 - \frac{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} e^{-nN_{OS1}} + \frac{1+e^{-N_{OS2}}}{1-e^{-N_{OS2}}} \frac{N_{OS2}}{2}(1-e^{-nN_{OS1}})}{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} e^{-nN_{OS1}} + (1-e^{-nN_{OS2}})}$
Standard SI	$E_{IS}^n = 1 - \frac{N_{OS1}}{1-e^{-N_{OS1}}} \frac{1-\left(\frac{1}{2-e^{-N_{OS1}}}\right)^n}{1-e^{-nN_{OS1}}}$	$E_{IS}^n = 1 - \frac{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} \frac{1}{(2-e^{-N_{OS2}})^n} + \frac{N_{OS2}}{1-e^{-N_{OS2}}}\left[1-\frac{1}{(2-e^{-N_{OS2}})^n}\right]}{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} e^{-nN_{OS1}} + (1-e^{-nN_{OS2}})}$
Extrapolating SI	$E_{IE}^n = 1 - \frac{\frac{N_{OS1}}{2}}{1-e^{-\frac{N_{OS1}}{2}}} \frac{1-\left(\frac{e^{-\frac{N_{OS1}}{2}}}{2-e^{-\frac{N_{OS1}}{2}}}\right)^n}{1-e^{-nN_{OS1}}}$	$E_{IE}^n = 1 - \frac{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} \left(\frac{e^{-\frac{N_{OS2}}{2}}}{2-e^{-\frac{N_{OS2}}{2}}}\right)^n + \frac{N_{OS2}}{1-e^{-\frac{N_{OS2}}{2}}}\left[1-\left(\frac{e^{-\frac{N_{OS2}}{2}}}{2-e^{-\frac{N_{OS2}}{2}}}\right)^n\right]}{\frac{k_T \Delta_t(kM^0)}{\Delta_t(k\Phi)} e^{-nN_{OS2}} + (1-e^{-nN_{OS2}})}$
Symmetric SI	$E_{IR}^n = 0$	$E_{IR}^n = 0$

equivalence between the time scale and the reaction rate. Indeed, it is clearly shown that Δt always appears with k or k_{tot} in mass balance equations.

3.7.3. Convergence rate of SI schemes

From the expression of the convergence criteria (Eqs. (22) and (34)), the iterative schemes, i.e. standard, extrapolating, and symmetric SIs, are compared, and the number of iterations required to obtain convergence for each SI scheme is estimated.

Using the mass balance expression at time step $(n+1)$ as a function of the iteration given by the standard SI scheme (Eq. (56)), the convergence criterion (Eq. (22)) for species i for a first-order reversible equilibrium can be calculated as:

$$CV_{IS}^{m} = \left| \frac{(-1)^{m-1}\left(1-e^{-k_{tot}\Delta t}\right)^{m-1}}{\frac{k_{tot}M_{IS_i}^{n}+\Phi_i k_{tot}\Delta t}{\Delta_t(kM_{IS}^{n})+\Delta_t(k\Phi)\Delta t}+\frac{1-e^{-k_{tot}\Delta t}}{2-e^{-k_{tot}\Delta t}}\left[1-(-1)^{m-1}\left(1-e^{-k_{tot}\Delta t}\right)^{m-1}\right]} \right| \tag{75}$$

In the case of an irreversible first-order reaction, the convergence criterion can be simplified to:

$$CV_{IS}^{m} = \frac{\left(1-e^{-k\Delta t}\right)^{m-1}}{1-(-1)^{m}\left(1-e^{-k\Delta t}\right)^{m}}\left(2-e^{-k\Delta t}\right) \tag{76}$$

The convergence criterion for the extrapolating SI scheme is the same with $\Delta t/2$ as for the standard SI scheme. Two iterations are needed to obtain convergence of the mass balance formulation for the symmetric SI scheme.

3.7.4. Efficiency of the schemes

To easily compare the schemes, the same relative mass balance error at steady state is imposed on each scheme and the computational effort is then estimated. In order to maintain generality, the computational effort is expressed in a specific cost unit (CU) per time unit. CU is defined as the cost of the computation of one transport plus one chemistry operator (Table 3). By imposing k or k_{tot} equal to 1, the relative mass balance errors at steady state

Table 3
Calculation of the computational effort for each scheme for a requested relative mass balance error of 1.00×10^{-3}

Scheme	Time step length (time unit)	Favourable case		Unfavourable case	
		Number of CUs[a] per time step	Computational effort: number of CUs per time unit	Number of CUs per time step	Computational effort: number of CUs per time unit
Standard SNI	$\Delta t = 2 \times 10^{-3}$	1	500	1	500
Strang-splitting SNI	$\Delta t = 0.1095$	1	9.13	2	18.3
Standard SI	$\Delta t = 2 \times 10^{-3}$	3	1500	5	2500
Extrapolating SI	$\Delta t = 4 \times 10^{-3}$	3	750	6	1500
Symmetric SI	Not limited	2	–	2	–

[a] One CU (cost unit): resolution of both one transport plus one chemistry operator.

(see Table 1) can be used to calculate the time step required by each scheme to reach the specified error. For each error level, the corresponding N_{OS} value is used to estimate the number of iterations needed for SI schemes (Eqs. (75) and (76)). The number of CUs per time unit can be then estimated for each scheme. However, the computational effort of the Strang-splitting SNI and the extrapolating SI schemes cannot be exactly expressed as number of CUs. The variation of the computational effort vs. the requested relative mass balance error is shown for the two following cases which frame the real effort (Fig. 1).

(i) The first one corresponds to a favourable case for which the computational effort per time step is minimal. A time step done by the Strang-splitting SNI scheme needs one CU. The effort for solving the transport operator is then negligible with respect to the computation of the chemistry one. For iterative schemes, the number of iterations requested for the used dimensionless number N_{OS} is minimal (Fig. 2). For the extrapolating SI scheme, the effort requested by the extrapolation procedure equals zero CU (Table 3).
(ii) For the unfavourable case, where the computational effort per time is maximal, a time step done by the Strang-splitting SNI scheme needs two CUs. The effort for solving

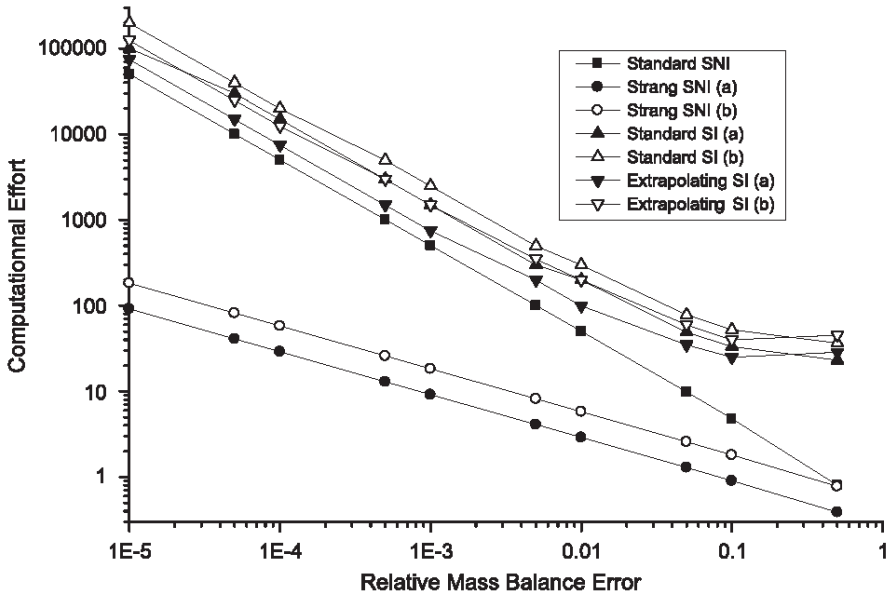


Fig. 1. Comparison of the computational effort (CU per time unit) vs. relative mass balance error for each scheme at steady state. One cost unit (CU) corresponds to the computation of one transport and one chemistry operator. (a) Favourable case: the computational effort is minimal at each time step. One CU per Strang-splitting SNI scheme time step: number of iterations used by standard and extrapolating SI schemes is minimal (see the (b) or (c) curves of Fig. 2). Extrapolation procedure for the extrapolating SI scheme needs zero CU. (b) Unfavourable case: the computational effort is maximal at each time step. Two CUs per Strang-splitting SNI scheme time step: number of iterations used by standard and extrapolating SI schemes is maximal (see the curve (a) of Fig. 2). Extrapolation procedure for the extrapolating SI scheme needs one CU.

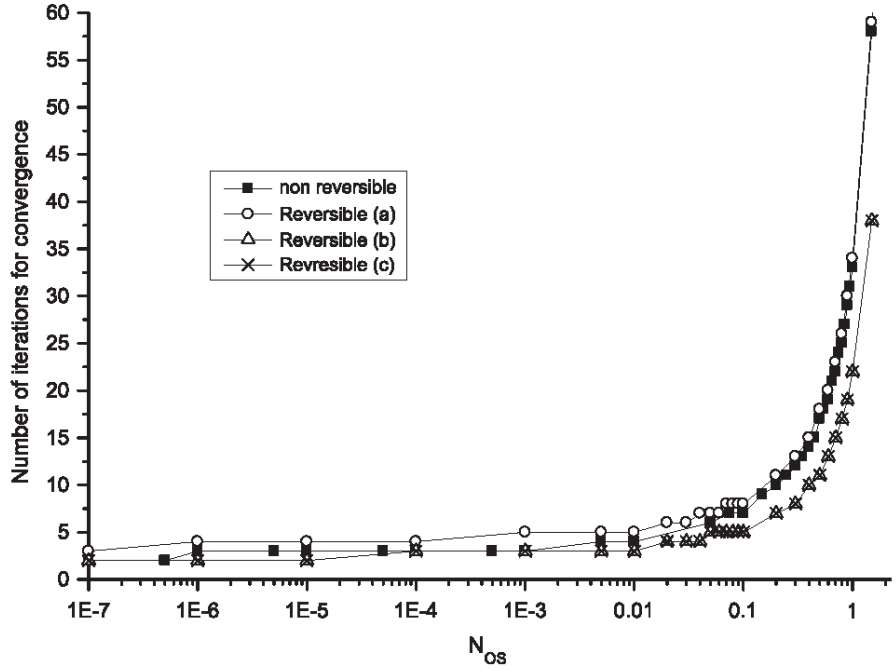


Fig. 2. Number of iterations needed to ensure convergence at $\varepsilon = 10^{-6}$ vs. N_{OS} for standard SI scheme with a first-order irreversible and reversible reaction. For reversible reaction: curves (a): $\frac{k_{tot}M^n_{IS_i} + \Phi_i k_{tot}\Delta t}{\Delta_i(kM^n_{IS}) + \Delta_i(k\Phi)\Delta t} = 0$; curves (b): $\frac{k_{tot}M^n_{IS_i} + \Phi_i k_{tot}\Delta t}{\Delta_i(kM^n_{IS}) + \Delta_i(k\Phi)\Delta t} = 100$; curves (c): $\frac{k_{tot}M^n_{IS_i} + \Phi_i k_{tot}\Delta t}{\Delta_i(kM^n_{IS}) + \Delta_i(k\Phi)\Delta t} = -100$.

the chemistry operator is then negligible with respect to the computation of the transport one. For iterative schemes, the number of iterations requested for the used dimensionless number N_{OS} is maximal (Fig. 2). For the extrapolating SI scheme, the effort requested by the extrapolation procedure equals one CU (Table 3).

4. Comparison of operator-splitting schemes

4.1. About transient boundary conditions

In order to simplify the calculation, the flux at the inlet of the domain is assumed to be constant over time (see Eq. (41)). But can this simplification have a significant effect on the results, and can the various schemes perform differently for transient boundary conditions? Even transient boundary conditions are described by modellers as constant boundary conditions over one time step. In this paper, results obtained for the first time step under a constant boundary conditions are then exact for transient boundary conditions at the first time step. Moreover, we show that the boundary conditions do not play any role in the relative mass balance error for irreversible reaction (see Table 2).

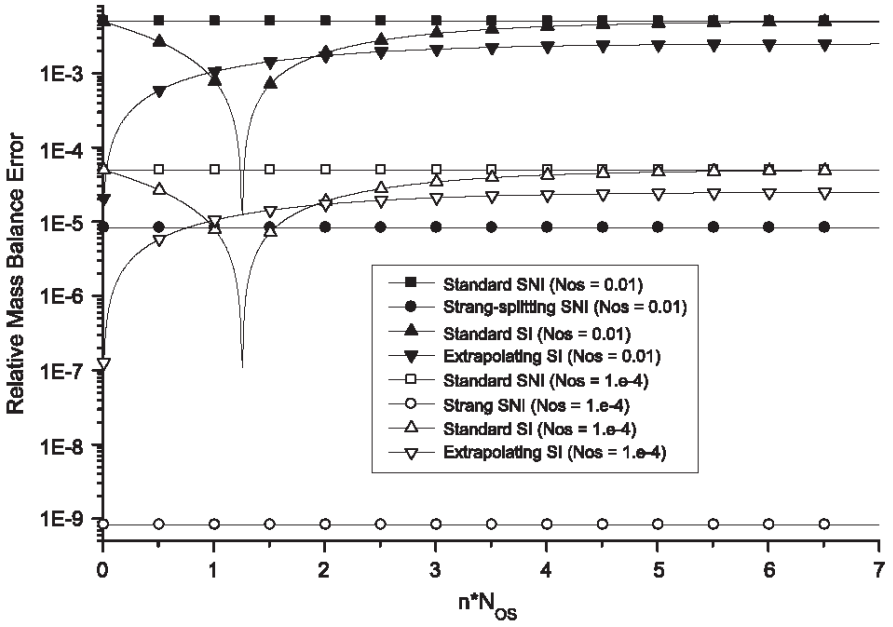


Fig. 3. Time dependence of the relative mass balance for a first-order irreversible reaction and a first-order reversible reaction if $\frac{k_{tot}\Delta_i(kM^0)}{\Delta_i(k\Phi)} = 0$, i.e. initial mass at thermodynamic equilibrium.

Figs. 3 and 4 show that errors at the first time steps and errors at steady state are very similar. So, our results can be extrapolated from constant to transient boundary conditions.

4.2. Mass balance errors at steady state

Mass balance errors at steady state are obtained after an infinite number of time steps (Table 1). For a first-order reaction rate, the evolution of relative mass balance errors vs. the dimensionless time is the same for one or two species (Table 1). Moreover, relative mass balance errors depend on dimensionless time exclusively. Taylor expansions of relative mass balance errors at steady state reported in Table 1 are also consistent with previous studies for SNI schemes (Hundsdorfer and Verwer, 1995; Valocchi and Malmstead, 1992; Kaluarachchi and Morshed, 1995a,b; Lanser and Verwer, 1998). Standard SNI is first order and Strang-splitting SNI scheme second order accurate. Standard and extrapolating SIs are first order accurate. The Standard SNI scheme provides a positive relative mass balance error and an underestimation of the mass balance (Table 1). For large values of N_{OS}, its relative mass balance error equals 1 as shown in Fig. 5. On the contrary, the Strang-splitting SNI, the standard and the extrapolating SIs schemes yield to negative relative masse balance errors (Table 1) what overestimates the mass balance. For small values of N_{OS}, the relative mass balance errors calculated from the standard SNI and SI schemes are equal in absolute value but

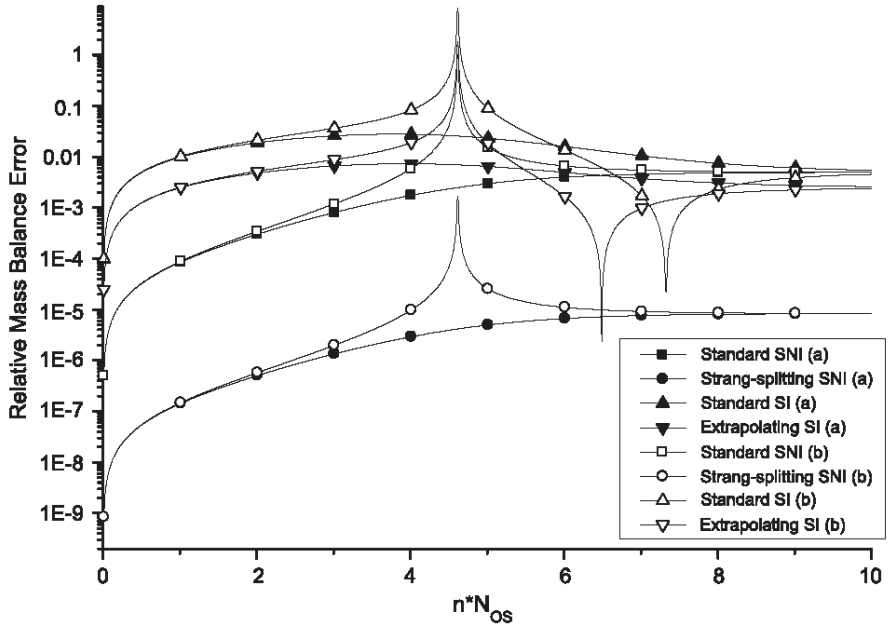


Fig. 4. Time dependence of the relative mass balance for a first-order reversible reaction: curves (a): $\frac{k_{tot}\Delta_i(kM^0)}{\Delta_i(k\Phi)} = 100$, initial mass and boundary flux at the same disequilibrium; curves (b): $\frac{k_{tot}\Delta_i(kM^0)}{\Delta_i(k\Phi)} = -100$, initial mass and boundary flux at inverse disequilibrium.

have opposite signs. For large values of N_{OS}, the relative mass balance errors supplied by the Strang-splitting SNI and the extrapolating SI schemes are the same as shown in Fig. 5.

Large values of the dimensionless time end up in unacceptable relative mass balance errors for all schemes (Fig. 5), except for the symmetric SI one. The Strang-splitting SNI scheme appears to be more accurate than the other iterative and non-iterative schemes, except for the symmetric SI one. Moreover, the accuracy given by the standard and the extrapolating SI schemes are very close to the one obtained from the standard SNI scheme.

4.3. Time dependence of relative mass balance errors

For the given conditions, i.e. initial mass, boundary flux, rate constant and time step length, the variation of the relative mass balance errors during a computation is studied. Relationships for first-order irreversible and reversible reactions are reported in Table 2.

For an irreversible first-order reaction, results of variation of the relative mass balance errors obtained at steady state can be extended to transient state (Fig. 3). The dimensionless time N_{OS1} is the unique parameter controlling the relative mass balance errors. Discontinuities in relative mass balance error variation for the standard SI scheme are due to change in the sign (Fig. 3). This scheme, which induces a mass balance

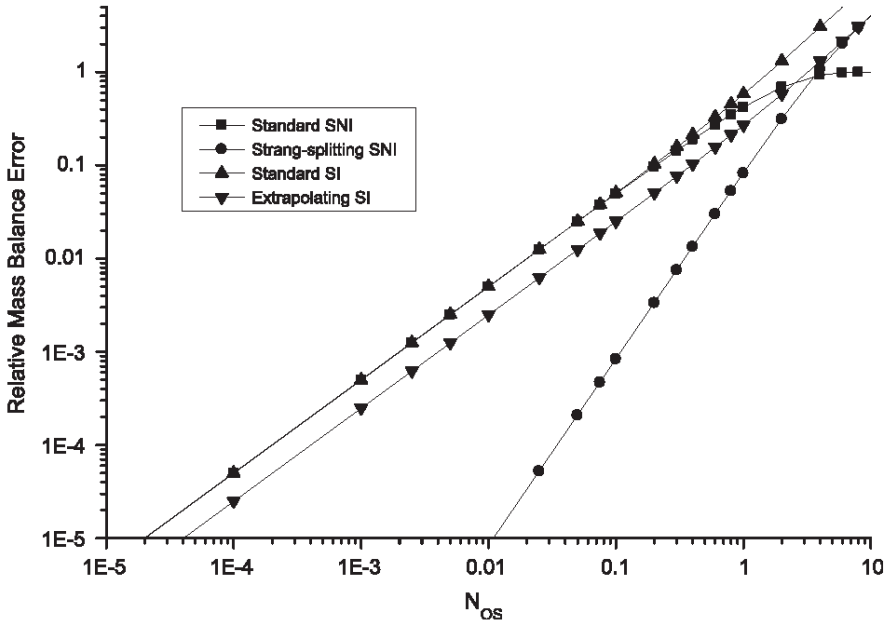


Fig. 5. Evolution of the relative mass balance errors at steady state vs. dimensionless times for different schemes and chemistry conditions. SI schemes iterate to infinity.

underestimation for the first time steps, provides an overestimation of the mass at steady state (Table 2). As N_{OS} are relatively small in Fig. 3, the absolute value of relative mass balance errors calculated from the standard SNI and SI schemes are equal at steady state. Whatever the number of time steps, the mass balance is underestimated by the standard SNI scheme, whereas it is overestimated by the Strang-splitting SNI and the extrapolating SI schemes.

For a first-order reversible reaction, the variation of the relative mass balance error depends on the dimensionless time N_{OS2} and $\frac{k_{tot}\Delta_i(kM^0)}{\Delta_i(k\Phi)}$. The latter one corresponds to the ratio of the deviation from equilibrium of the initial mass and of the boundary flux for the formation of species i. If the initial mass is zero or if the system is initially at equilibrium, the relative mass balance errors of a reversible or a irreversible reaction are the same (see Fig. 3). The influence of $\frac{k_{tot}\Delta_i(kM^0)}{\Delta_i(k\Phi)}$ on the variation of the relative mass balance error is reported in Fig. 4. The behaviour of the reversible reaction is globally the same as for the irreversible one, whatever the initial mass and boundary flux disequilibrium is. We observe discontinuities in the variation of relative mass balance errors due to a change in sign (see Fig. 3). Evolutions of the mass balance estimation are similar for reversible and irreversible case: (i) the standard SNI scheme underestimates the mass balance; (ii) the Strang-splitting SNI and the extrapolating SI schemes overestimate it; and (iii) the standard SI underestimates for first time steps and overestimates the mass balance at steady state.

4.4. Convergence rate of SI schemes

The convergence criteria for a first-order irreversible reaction depend only on the dimensionless number N_{OS1} (see Eq. (76)). For a first-order reversible reaction, the dimensionless number N_{OS2} is the main parameter in Eq. (75) associated with the expression $\frac{k_{tot}M_{IS_i}^n+\Phi_i k_{tot}\Delta t}{\Delta_i(kM_{IS}^n)+\Delta_i(k\Phi)\Delta t}$. The value of $\frac{k_{tot}M_{IS_i}^n+\Phi_i k_{tot}\Delta t}{\Delta_i(kM_{IS}^n)+\Delta_i(k\Phi)\Delta t}$, which changes during the computation of a reactive transport problem because it depends on the time step n, does not have a significant influence on the number of iterations needed to converge as shown in Fig. 2. In this case of the standard and the extrapolating SI schemes, the number of iterations needed to converge is thus quite constant during the computation. Moreover, for $N_{OS} < 0.01$, the convergence rate for the standard and the extrapolating SI schemes does not vary too much. The convergence is very fast, taking less than five iterations. If $N_{OS}>0.01$, the number of iterations increases very quickly in N_{OS}.

4.5. Efficiency of the schemes

A first-order relationship between the computational effort and the relative mass balance errors is observed for the standard SNI, the standard SI and the extrapolating SI schemes (Fig. 1). A second-order relationship between the computational effort and the relative mass balance errors for both favourable and unfavourable cases is observed for the Strang-splitting SNI scheme. For $N_{OS}>0.1$ and for the standard and the extrapolating SI schemes, the computational effort does not decrease more even if N_{OS} grows (Fig. 1). In this case, the decrease in the number of time steps, which is due to the increase in Δt, is associated with an increase in the number of iterations needed per time step (see Fig. 2). The combination of both effects limits the decrease in the computational effort. From the reported difficulties associated with the implementation of iterative OS schemes (Steefel and MacQuarrie, 1996) and from our results, the standard and the extrapolating SI schemes should not be used. We shall prefer the symmetric SI or at least the SNI schemes, especially the Strang-splitting SNI scheme, which provides the most precise results without convergence or stability problems.

Whatever the favourable or unfavourable case, the results of this comparison are the same: the symmetric SI scheme is the most accurate one. The Strang-splitting SNI is more efficient than the standard SNI, standard SI and extrapolating SI schemes.

5. Conclusion

Results obtained with several non-iterative and iterative operator-splitting schemes are compared. The benefit of working with mass balances rather than with concentrations comes from the ability to exactly solve each operator, i.e. transport, chemistry or reactive transport for any initial condition. Therefore, the estimated errors do not depend on the technique used to solve each operator, but on the splitting method itself. The results of this work can be applied independently of the

methods used to solve the transport and chemistry equations. Irreversible and reversible first-order chemical reactions are studied and provide similar conclusions. We show the existence of dimensionless numbers, $N_{OS1}=k\Delta t$ and $N_{OS2}=k_{tot}\Delta t$, which control the operator-splitting error. Irreversible or reversible chemical reaction leads to the same operator-splitting error. Since the time step appears always with N_{OS} in error relationships, N_{OS} is considered as a controlling parameter rather than Δt.

With respect to the efficiency, the following order (from the most to the less efficient scheme) is proposed: symmetric SI, Strang-splitting SNI, standard SNI, extrapolating SI, and standard SI.

These results are consistent with the case of irreversible and reversible chemical reactions. The symmetric SI scheme does not induce any operator-splitting errors. The Strang-splitting SNI is $O(N_{OS}^2)$ accurate and other schemes (standard SNI, standard SI and extrapolating SI) are first order accurate.

It must be underlined that these results are obtained with a kinetic formulation of chemical phenomena. Extrapolating these results to an instantaneous equilibrium formulation must be avoided, since the assumptions underlying this formulation are incompatible with those of the kinetic formulation. In fact, OS errors depend on the intrinsic mathematical nature of the operators which are split (Carrayrou, 2001). Results obtained for ordinary differential equations (kinetic formulation) cannot be applied to nonlinear algebraic equations (equilibrium formulation).

In this study, the limit proposed by Valocchi and Malmstead (1992) for the standard and the alternating SNI schemes and for irreversible reactions is extended to the standard SNI, the Strang-splitting SNI, the standard SI and the symmetric SI schemes with reversible and irreversible first-order reactions: N_{OS} must be kept <0.1 to obtain a mass balance error $<1\%$ for all schemes at steady state. The number of iterations necessary to converge with the standard SI and the extrapolating SI schemes increases rapidly when $N_{OS}>0.1$. This limit can be used for standard SNI, Strang-splitting SNI, standard SI and extrapolating SI schemes and for irreversible and reversible first-order kinetics.

The above results are obtained for linear reversible and irreversible reactions, which are representative of many chemical reactions. These results give useful preliminary statements for implementing OS method into reactive transport models for more complex problems.

A large range of kinetic constant values has been observed. The symmetric SI scheme appears to be the only one, which does not impose the time step required for the fastest kinetics. The symmetric SI scheme is thus very attractive. The exact solution is obtained at each time step, independently on N_{OS}. Further study over OS method in reactive transport with kinetic formulation of the chemical operator would focus on the symmetric SI scheme, which is very efficient for reducing time-dependent OS errors. This scheme would be used and would be coupled to reactive transport model giving a reduction of the space-dependent error (numerical diffusion) such as numerical correction of diffusion, corrected boundary (Hundsdorfer and Verwer, 1995) or reduction of numerical diffusion by discontinuous finite elements (Siegel et al., 1997).

List of Notations

Notation	Signification [Dimension]
k	Kinetic constant [T^{-1}]
x	Abscissa [L]
t	Time [T]
u	Flux velocity [$L\ T^{-1}$]
$\boldsymbol{D}$	Dispersion tensor [$L^2\ T^{-1}$]
c	Aqueous concentration [$M\ L^{-3}$]
c_0	Aqueous concentration at boundary [$M\ L^{-3}$]
$\boldsymbol{C}$	Vector of aqueous concentration at the nodes of the domain [$M\ L^{-3}$]
Φ	Imposed flux at boundary [$M\ T^{-1}$]
M	Mass present in the domain [M]
M_T	Mass present in the domain calculated by the transport operator [M]
M_C	Mass present in the domain calculated by the chemistry operator [M]
E	Error on the mass balance [−]
$\Delta_i(kX)$	Deviation from equilibrium of variable X [$X\ T^{-1}$]
ε	Convergence limit [−]
CV_i	Convergence criterion for species i [−]
R_C	Chemistry source–sink term [$M\ T^{-1}$]
R_T	Transport source–sink term [$M\ T^{-1}$]
$\boldsymbol{R}_C$	Vector of the chemistry source–sink term at the nodes of the domain [$M\ T^{-1}$]
$\boldsymbol{R}_T$	Vector of the transport source–sink term at the nodes of the domain [$M\ T^{-1}$]
$(\cdot)^n$	Index of the time step
$(\cdot)^m$	Index of iteration
$(\cdot)_i$	Index of the species
$(\cdot)_{EX}$	Index of the exact resolution
$(\cdot)_{SS}$	Index of the standard SNI scheme
$(\cdot)_{ST}$	Index of the Strang-splitting SNI scheme
$(\cdot)_{IT}$	Index of the standard SI scheme
$(\cdot)_{IE}$	Index of the extrapolating SI scheme
$(\cdot)_{IR}$	Index of the symmetric SI scheme

The dimensions are given for a mono-directional resolution of the reactive transport problem. [L] is the length dimension, [M] the mass dimension, [T] the time dimension, and [−] means no dimension.

6. Uncited reference

Liu and Narasimhan, 1989

Acknowledgements

We thank Alexandre Ern from the "Ecole Nationale Supérieure des Ponts et Chaussées" for his helpful comments, and the referees and the editor for their useful

suggestions. J.C. has been supported by a grant from the "Ministère de l'Education Nationale, de la Recherche et de la Technologie". This work was supported by the "Programme Environnement, Vie and Sociétés" of CNRS.

Appendix A. Standard SNI scheme

Multiplying the recursive form (Eq. (47)) by k_i for $i=1$, 2, and subtracting these two equations yields:

$$\Delta_i(kM_{SS}^{n+1}) = \Delta_i(kM_{SS}^{n}) + \Delta_i(k\Phi)\Delta t - [\Delta_i(kM_{SS}^{n}) + \Delta_i(k\Phi)\Delta t](1 - e^{-k_{tot}\Delta t}) \quad (A1)$$

Simplifying Eq. (A1) provides a recursive expression of the $\Delta_i(kM_{SS}^{n})$ series:

$$\Delta_i(kM_{SS}^{n+1}) = \Delta_i(kM_{SS}^{n})e^{-k_{tot}\Delta t} + \Delta_i(k\Phi)\Delta t\, e^{-k_{tot}\Delta t} \quad (A2)$$

Eq. (A2) can be explicitly formulated for $n \geq 1$:

$$\Delta_i(kM_{SS}^{n}) = \Delta_i(kM_{SS}^{0})e^{-nk_{tot}\Delta t} + \Delta_i(k\Phi)\Delta t \sum_{i=1}^{n} e^{-ik_{tot}\Delta t} \quad (A3)$$

After simplification, this equation becomes:

$$\Delta_i(kM_{SS}^{n}) = \Delta_i(kM_{SS}^{0})e^{-nk_{tot}\Delta t} + \Delta_i(k\Phi)\Delta t \frac{e^{-k_{tot}\Delta t} - e^{-(n+1)k_{tot}\Delta t}}{1 - e^{-k_{tot}\Delta t}} \quad (A4)$$

Incorporating Eq. (A4) into Eq. (47) gives a recursive formulation of the $M_{SS_i}^{n}$ series:

$$M_{SS_i}^{n+1} = M_{SS_i}^{n} + \left(\Phi_i + \frac{\Delta_i(k\Phi)}{k_{tot}}\right)\Delta t + \frac{\Delta_i(kM_{SS}^{0})(1 - e^{-k_{tot}\Delta t}) - \Delta_i(k\Phi)\Delta t\, e^{-k_{tot}\Delta t}}{k_{tot}} \times e^{-nk_{tot}\Delta t} \quad (A5)$$

Eq. (A5) can be explicitly written as:

$$M_{SS_i}^{n} = M_{SS_i}^{0} + \left(\Phi_i + \frac{\Delta_i(k\Phi)}{k_{tot}}\right)n\Delta t + \frac{\Delta_i(kM_{SS}^{0})(1 - e^{-k_{tot}\Delta t}) - \Delta_i(k\Phi)\Delta t\, e^{-k_{tot}\Delta t}}{k_{tot}} \times \sum_{i=0}^{n-1} e^{-ik_{tot}\Delta t} \quad (A6)$$

Rearranging Eq. (A6) gives Eq. (48).

Appendix B. Strang-splitting SNI scheme

Multiplying the recursive form (Eq. (50)) by k_i for $i=1$ and 2, and subtracting these two equations gives a recursive expression of the $\Delta_i(kM_{ST}^n)$ series:

$$\Delta_i(kM_{ST}^{n+1}) = \Delta_i(kM_{ST}^n)e^{-k_{tot}\Delta t} + \Delta_i(k\Phi)\frac{\Delta t}{2}(1+e^{-k_{tot}\Delta t}) \tag{B1}$$

An explicit formulation of this series is obtained as:

$$\Delta_i(kM_{ST}^n) = \Delta_i(kM_{ST}^0)e^{-nk_{tot}\Delta t} + \Delta_i(k\Phi)\frac{\Delta t}{2}(1+e^{-k_{tot}\Delta t})\sum_{i-1}^{n-1} e^{-ik_{tot}\Delta t} \tag{B2}$$

After simplification, the result is:

$$\Delta_i(kM_{ST}^n) = \Delta_i(kM_{ST}^0)e^{-nk_{tot}\Delta t} + \Delta_i(k\Phi)\frac{\Delta t}{2}\frac{1+e^{-k_{tot}\Delta t}}{1-e^{-k_{tot}\Delta t}}(1-e^{-nk_{tot}\Delta t}) \tag{B3}$$

Incorporating Eq. (B3) into Eq. (50) provides a new formulation of the $\Delta_i(kM_{ST}^n)$ series:

$$M_{ST_i}^{n+1} = M_{ST_i}^n + \left(\Phi_i + \frac{\Delta_i(k\Phi)}{k_{tot}}\right)\Delta t + \left[\frac{\Delta_i(kM_{ST}^n)}{k_{tot}}(1-e^{-k_{tot}\Delta t}) + \frac{\Delta_i(k\Phi)}{k_{tot}}\frac{\Delta t}{2}(1-e^{-k_{tot}\Delta t})\right] \times e^{-nk_{tot}\Delta t} \tag{B4}$$

An explicit formulation of Eq. (B4) is:

$$M_{ST_i}^n = M_{ST_i}^0 + \left(\Phi_i + \frac{\Delta_i(k\Phi)}{k_{tot}}\right)n\Delta t + \left[\frac{\Delta_i(kM_{ST}^0)}{k_{tot}}(1-e^{-k_{tot}\Delta t}) + \frac{\Delta_i(k\Phi)}{k_{tot}}\frac{\Delta t}{2}(1-e^{-k_{tot}\Delta t})\right] \times \sum_{i-0}^{n-1} e^{-ik_{tot}\Delta t} \tag{B5}$$

After simplification of Eq. (B5), the solution (Eq. (51)) is obtained.

Appendix C. Standard SI scheme

The recursive formulation of the $\Delta_i(kM_T^m[(n+1)\Delta t])$ series is:

$$\Delta_i(kM_T^{m+1}[(n+1)\Delta t]) = \Delta_i(kM_{IS}^n) + \Delta_i(k\Phi)\Delta t - \Delta_i(kM_T^{m+1}[(n+1)\Delta t])(1-e^{-k_{tot}\Delta t}) \tag{C1}$$

An explicit form of this series is:

$$\Delta_i(kM_T^m[(n+1)\Delta t]) = [\Delta_i(kM_{IS}^n) + \Delta_i(k\Phi)\Delta t]\sum_{i-0}^{m-1}(-1)^i(1-e^{-k_{tot}\Delta t})^i \tag{C2}$$

Simplifications yield:

$$\Delta_i(kM_{\mathrm{T}}^m[(n+1)\Delta t]) = [\Delta_i(kM_{\mathrm{IS}}^n) + \Delta_i(k\Phi)\Delta t]\frac{1-(-1)^m(1-e^{-k_{\mathrm{tot}}\Delta t})^m}{2-e^{-k_{\mathrm{tot}}\Delta t}} \tag{C3}$$

Incorporating Eq. (C3) into Eq. (55) yields Eq. (56).

Eq. (57) written for $i=1$ and 2 provides the recursive form of the $\Delta_i(kM_{\mathrm{IS}}^n)$ series:

$$\Delta_i(kM_{\mathrm{IS}}^{n+1}) = \frac{\Delta_i(kM_{\mathrm{IS}}^n) + \Delta_i(k\Phi)\Delta t}{2-e^{-k_{\mathrm{tot}}\Delta t}} \tag{C4}$$

An explicit form of this series is:

$$\Delta_i(kM_{\mathrm{IS}}^n) = \Delta_i(kM_{\mathrm{IS}}^0)\left(\frac{1}{2-e^{-k_{\mathrm{tot}}\Delta t}}\right)^n + \Delta_i(k\Phi)\Delta t\sum_{i=1}^{n}\left(\frac{1}{2-e^{-k_{\mathrm{tot}}\Delta t}}\right)^i \tag{C5}$$

Eq. (C5) can be simplified to give:

$$\Delta_i(kM_{\mathrm{IS}}^n) = \Delta_i(kM_{\mathrm{IS}}^0)\left(\frac{1}{2-e^{-k_{\mathrm{tot}}\Delta t}}\right)^n + \frac{\Delta_i(k\Phi)\Delta t}{1-e^{-k_{\mathrm{tot}}\Delta t}}\left[1-\left(\frac{1}{2-e^{-k_{\mathrm{tot}}\Delta t}}\right)^n\right] \tag{C6}$$

After incorporating Eq. (C6) into Eq. (57), a recursive form of the mass balance series is obtained:

$$M_{\mathrm{IS}_i}^{n+1} = M_{\mathrm{IS}_i}^n + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right)k_{\mathrm{tot}}\Delta t + \left[\frac{\Delta_i(kM_{\mathrm{IS}}^0)(1-e^{-k_{\mathrm{tot}}\Delta t})}{k_{\mathrm{tot}}} - \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}k_{\mathrm{tot}}\Delta t\right] \times \left(\frac{1}{2-e^{-k_{\mathrm{tot}}\Delta t}}\right)^{n+1} \tag{C7}$$

An explicit form of Eq. (C7) is:

$$M_{\mathrm{IS}_i}^n = M_{\mathrm{IS}_i}^0 + \left(\frac{\Phi_i}{k_{\mathrm{tot}}} + \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}\right)nk_{\mathrm{tot}}\Delta t + \left[\frac{\Delta_i(kM_{\mathrm{IS}}^0)(1-e^{-k_{\mathrm{tot}}\Delta t})}{k_{\mathrm{tot}}} - \frac{\Delta_i(k\Phi)}{k_{\mathrm{tot}}^2}k_{\mathrm{tot}}\Delta t\right] \times \sum_{i=1}^{n}\left(\frac{1}{2-e^{-k_{\mathrm{tot}}\Delta t}}\right)^i \tag{C8}$$

Simplification of Eq. (C8) leads to the solution (Eq. (58)).

References

Appelo, C.A.J., Verweij, E., Schäfer, H., 1997. A hydrogeochemical transport model for an oxidation experiment with pyrite/calcite/exchangers/organic matter containing sand. Appl. Geochem. 13, 257–268.

Barry, D.A., Miller, C.T., Culligan-Hensley, P.J., 1996. Temporal discretisation errors in non-iterative split-operator approaches to solving chemical reaction/groundwater transport models. J. Contam. Hydrol. 22, 1–17.

Barry, D.A., Miller, C.T., Culligan, P.J., Bajracharya, K., 1997. Analysis of split operator methods for nonlinear and multispecies groundwater chemical transport models. Math. Comput. Simul. 43, 331–341.

Carrayrou, J., 2001. Modélisation du transport de solutés réactifs en milieu poreux. PhD thesis, Louis Pasteur University, Strasbourg.

Cederberg, A., Street, R.L., Leckie, J.O., 1985. A groundwater mass transport and equilibrium chemistry model for multicomponent systems. Water Resour. Res. 21, 1095–1104.

Ginn, T.R., 2001. Stochastic-convective transport with nonlinear reactions and mixing: finite streamtube ensemble formulation for multicomponent reaction systems with intra-streamtube dispersion. J. Contam. Hydrol. 47, 1–28.

Herzer, J., 1989. CHEMFLO-Dokumentation eines Schadstoffmodells für mehrere wechselwirkende Komponenten im Grundwasser. Institut für Wasserbau, Universität Stuttgart, Germany. Bericht Nr. 89/34 (HG 118).

Hesstvedt, E., Hov, O., Isaksen, I.S.A., 1978. Quasi-steady-state approximations in air pollution modeling: comparison of two numerical schemes for oxidant prediction. Int. J. Chem. Kinet. 10, 971–994.

Hundsdorfer, W.H., 1996. Numerical solution of advection–diffusion-reaction equations. Report NM-R9603, CWI, Amsterdam.

Hundsdorfer, W., Verwer, J.G., 1995. A note on splitting errors for advection-reaction equations. Appl. Numer. Math. 18, 191–199.

Jauzein, M., André, C., Margrita, R., Sardin, M., Schweich, D., 1989. A flexible computer code for modelling transport in porous media: IMPACT. Geoderma 44, 95–113.

Kaluarachchi, J.J., Morshed, J., 1995a. Critical assessment of the operator-splitting technique in solving the advection–dispersion-reaction equation: 1. First-order reaction. Adv. Water Resour. 18, 89–100.

Kaluarachchi, J.J., Morshed, J., 1995b. Critical assessment of the operator-splitting technique in solving the advection–dispersion-reaction equation: 2. Monod kinetics and coupled transport. Adv. Water Resour. 18, 101–110.

Kanney, J.F., Miller, C.T., Kelley, C.T., 2003. Convergence of iterative split-operator for approximating nonlinear reactive transport problem. Adv. Water Resour. 26, 247–261.

Lanser, D., Verwer, J.G., 1998. Analysis of operator-splitting for advection–diffusion-reaction problems from air pollution modelling. Report MAS-R9805, CWI, Amsterdam.

Leeming, G.J.S., Mayer, K.U., Simpson, R.B., 1998. Effects of chemical reactions on iterative methods for implicit time stepping. Adv. Water Resour. 22, 333–347.

Liu, C.W., Narasimhan, T.N., 1989. Redox-controlled multiple-species reactive chemical transport: 1. Model development. Water Resour. Res. 25, 869–882.

McRae, G.J., Goodin, W.R., Seinfeld, J.H., 1982. Numerical solution of the atmospheric diffusion equation for chemically reacting flows. J. Comput. Phys. 45, 1–42.

Sandu, A., Verwer, J.G., Blom, J.G., Spee, E.J., Carmichael, G.R., Potra, F.A., 1997. Benchmarking stiff ODE solvers for atmospheric chemistry problems: II. Rosenbrock solvers. Atmos. Environ. 31, 3459–3472.

Siegel, P., Mosé, R., Ackerer, Ph., Jaffre, J., 1997. Solution of the advection-diffusion equation using a combination of discontinuous and mixed finite elements. Int. J. Numer. Methods Fluids 24, 595–613.

Šimůnek, J., Suarez, D.L., 1994. Two-dimensional transport model for variably saturated porous media with major ion chemistry. Water Resour. Res. 30, 1115–1133.

Steefel, C.I., MacQuarrie, K.T.B., 1996. Approaches to modelling of reactive transport in porous media. In: Lichtner, P.C., Steefel, C.I., Oelkers, E.H. (Eds.), Reactive Transport in Porous Media. Reviews in Mineralogy, vol. 34. Mineralogical Society of America, Washington, pp. 82–129.

Strang, G., 1968. On the construction and comparison of different schemes. SIAM J. Numer. Anal. 5, 506–517.

Valocchi, A.J., Malmstead, M., 1992. Accuracy of operator-splitting for advection–dispersion-reaction problems. Water Resour. Res. 28, 1471–1476.

Walter, L., Frind, E.O., Blowes, D.W., Ptacek, C.J., Molson, J.W., 1994. Modeling of multicomponent reactive transport in groundwater: 1. Model development and evaluation. Water Resour. Res. 30, 3137–3148.

Yeh, G.T., Tripathi, V.S., 1989. A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components. Water Resour. Res. 25, 93–108.

Zysset, A., 1993. Modellierung des chemischen Zustandes in Grundwasser-Infiltrations-System. PhD Thesis, Swiss Federal Institute of Technology (ETH), Zürich.

Zysset, A., Stauffer, F., Dracos, T., 1994. Modeling of chemically reactive groundwater transport. Water Resour. Res. 30, 2217–2228.

Annexe 3. New efficient algorithm for solving thermodynamic chemistry

New Efficient Algorithm for Solving Thermodynamic Chemistry

Jérôme Carrayrou, Robert Mosé, and Philippe Behra
Institut de Mécanique des Fluides et des Solides de l'Université Louis Pasteur, UMR 7507 ULP-CNRS, 67000 Strasbourg, France

Modeling thermodynamic equilibrium of complex nonlinear chemical systems with the most used Newton-Raphson method can lead to nonconvergence. From the mathematical properties of the set of equations, a chemically permitted interval is defined. By imposing this interval, the robustness of the Newton-Raphson method is increased at a low computing time cost. The new method, the positive continuous fraction, does not depend on the first derivative of the objective function to find the solution. A new algorithm is thus built by the association of this very robust method with the fast Newton-Raphson method and the definition of the chemically allowed interval. This combined algorithm is very impressive in terms of reliability, robustness, and speed.

Introduction

Chromatographic prediction and optimization involving multicomponent systems in chemical engineering or risk assessment in highly sensitive environmental domains such as heavy metal contamination (Chilakapati, 1999), nuclear waste disposal (White et al., 1984; Walsh et al., 1984), or nuclear testing (Bellot et al., 1999; Kersting et al., 1999) requires more and more calculation of reactive transport. The operator-splitting approach has mainly been used in reactive transport models by coupling mass transport equation and batch chemistry (Bryant et al., 1986; Krebs et al., 1987). In this case, the code solving the chemistry is called at least once per node and per time step.

In thermodynamic terms, a chemical equilibrium calculation, which attempts to find the minimum value for the Gibbs free energy, can be carried out in one of two ways: by minimizing a free energy function or by solving a set of nonlinear equations consisting of equilibrium constants and mass balance constraints. The two methods are thermodynamically equivalent, but the major disadvantage of using a free energy database is that these values are not nearly as reliable as directly measured equilibrium constants (Nordstrom and Ball, 1984). Many mathematical methods have been tested to solve the set of nonlinear algebraic equations describing thermodynamic equilibria. Zero-order methods such as the continuous fractions method (Wigley, 1977) and the Simplex method (Nelder and Mead, 1965; Wood, 1993) do not use the derivative of the objective function. The latter methods converge more slowly (Morin, 1985), but are sometimes considered more robust than first-order methods. The Simplex method is believed to be the most robust and may find the thermodynamic equilibrium when first-order methods are inefficient (Brassard and Bodurtha, 2000; Parkhurst and Appelo, 1999). First-order methods use the derivative of the objective function. The Newton-Raphson method is used the most to compute thermodynamic equilibrium in software such as MICROQL (Westall, 1979), IMPACT (Krebs et al., 1987), CHESS (van der Lee, 1998a,b), or PHREEQC (Parkhurst and Appelo, 1999a,b). Robustness (convergence for many cases), flexibility (easily modeling all chemical processes), and quickness (quadratic convergence (Morin, 1985)) are its great advantages. Nevertheless, the Newton-Raphson method is subject to nonconvergence for some ill-conditioned systems (van der Lee, 1998), and Brassard and Bodurtha (2000) have recently shown that this observation can be extended to other methods. Because of its frequent use, the Newton-Raphson method will be considered in this work as a reference method. Other methods such as the Gauss-Seidel, Gauss-Newton, or Levenberg-Marcard (Brassard and Bodurtha, 2000) can be used for calculating thermodynamic equilibrium. The Newton-Raphson method is sometimes associated with a relaxation technique (van der Lee, 1998) which controls the size of

Correspondence concerning this article should be addressed to Ph. Behra at this current address: Ecole Nationale Supérieure des Ingénieurs en Arts Chimiques et Technologiques, Laboratoire de Chimie Agro-Industrielle, UMR 1010 INRA/INP-ENSIACET, 118 route de Narbonne, 31077 Toulouse Cedex, 4-France.
Current address of R. Mosé: Engees, 1 quai koch, 67000 Strasbourg, France.

the method's down step and strengthens its robustness. This association can be done numerous ways with many relaxation techniques. Because of the high number of relaxation techniques or other methods applied to specific problems, these implementations will not be addressed in this work.

By assuming instantaneous equilibrium, it is known that chemistry computer codes are subject to nonconvergence (Walsh et al., 1984) which involves the loss of the computational effort or induce the wrong output. An accurate speciation code has to pursue the following criteria: reliability (exact solution), robustness (convergence certain), and quickness (computation time). The objective of this article is to present a new method, the Positive Continuous Fraction method, present a new constraint, respecting the chemically allowed interval (CAI), and propose a new algorithm, which respects these three criteria, to evaluate the thermodynamic equilibrium of nonlinear complex chemical systems. The principles reported here for examples of thermodynamic calculations are of great interest for every scientific domain that requires resolution of nonlinear algebraic systems.

Modeling

One very efficient formulation for the computation of thermodynamic equilibrium is based on the Tableau concept (referred to as Morel's Tableau) (Morel and Morgan, 1972; Morel, 1983). N_X components X_j are chosen among the N_C species C_i in order to write the formation of each species as a combination of the components. The mass action law for the formation of the C_i species is written with the equilibrium constant K_i and the stoichiometric coefficients $a_{i,k}$ for each component X_k

$$\{C_i\} = K_i \cdot \prod_{k=1}^{N_X} \{X_k\}^{a_{i,k}} \tag{1}$$

where $\{C_i\}$ and $\{X_k\}$ are the activities of species C_i and component X_k.

If N_{CP} precipitated species Cp_i are taken into account, the mass action law for the precipitation of Cp_i is written with the precipitation constant Kp_i such that $Kp_i = 1/Ks_i$, where Ks_i is the solubility product, and $ap_{i,k}$ is the stoichiometric coefficients. The saturation index (SI_i) of Cp_i is equal to its activity which is for a pure solid phase

$$SI_i = Kp_i \cdot \prod_{k=1}^{N_X} \{X_k\}^{ap_{i,k}} = 1 \tag{2}$$

The conservation of the total concentration $[T_j]$ of the jth component in the system is then written

$$[T_j] = \sum_{i=1}^{N_C} a_{i,j} \cdot [C_i] + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i] \tag{3}$$

where $[C_i]$ is the concentration of species C_i and $[Cp_i]$ is the amount of precipitated species Cp_i per liquid volume unit.

The activity of species C_i (respectively, component X_j) is calculated by the following relation

$$\{C_i\} = \gamma_i \cdot \frac{[C_i]}{[C^\circ]} \text{ and } \{X_j\} = \gamma_j \cdot \frac{[X_j]}{[C^\circ]} \tag{4}$$

where γ_i (respectively, γ_j) is the activity coefficient of species C_i (respectively, component X_j) and $[C^\circ]$ is the reference concentration: one mol per liter.

By substituting the mass action law (Eq. 1) in the mass conservation equation (Eq. 3), the following relationship, which depends only on the components and the precipitated species concentrations, is thus obtained

$$[T_j] = \sum_{i=1}^{N_C} a_{i,j} \cdot \left(\frac{K_i}{\gamma_i} \cdot \prod_{k=1}^{N_X} (\gamma_k \cdot [X_k])^{a_{i,k}} \right) + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i] \tag{5}$$

There are two ways by which this set of equations can be solved. The most commonly used consists of changing the set of components by taking the precipitated species as a new component into account instead of the nonprecipitated species (Westall, 1979). This method implies new calculations of the stoichiometric matrix $a_{i,j}$ and $ap_{i,j}$ and total concentration $[T_j]$. In order to minimize calculation time, we prefer to add both an unknown, the precipitated species amount, and an equation, its saturation index which equals one. It is a set of $(N_X + N_{CP})$ nonlinear algebraic equations which can be numerically solved through iterative methods. The values of the component $[X_k]$ and the precipitated species $[Cp_i]$ concentrations at equilibrium are then found when the $(N_X + N_{CP})$ objective functions Y_i equal zero

$$Y_j = -[T_j] + \sum_{i=1}^{N_C} a_{i,j} \left(\frac{K_i}{\gamma_i} \cdot \prod_{k=1}^{N_X} (\gamma_k \cdot [X_k])^{a_{i,k}} \right) + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i] \quad \text{for } j = 1 \text{ to } N_X$$

$$Y_{j-N_X-i} = 1 - Kp_i \cdot \prod_{k=1}^{N_X} (\gamma_k \cdot [X_k])^{ap_{i,k}} \quad \text{for } i = 1 \text{ to } N_{CP} \tag{6}$$

If some species can precipitate, a first equilibrium, called transitory equilibrium, is searched without taking any precipitated species into account ($N_{CP} = 0$) in Eq. 5. The saturation index is then calculated. If all the SIs are smaller than one, all the precipitated species are undersaturated and the final equilibrium is obtained at a first stage. If some species (chemical forms) are supersaturated, the SIs are greater than one and the most supersaturated chemical form is then taken into account ($N_{CP} = N_{CP} + 1$) in Eq. 5 for searching a new equilibrium. This procedure is repeated until all the SIs are equal to or smaller than one.

In the case of an ideal system assumption, the activity coefficients γ equal one. If an activity correction is required, approximations such as the Davies can be used for calculating activity coefficients (Stumm and Morgan, 1995). The equa-

tions are

$$\ln(\gamma_i) = -A \cdot z_i^2 \cdot \left(\frac{\sqrt{I}}{1+\sqrt{I}} - b \cdot I \right) \tag{7}$$

with

$$I = \frac{1}{2} \cdot \sum_i [C_i] \cdot z_i^2 \tag{8}$$

and

$$A = 1.82 \cdot 10^6 (\epsilon_w T)^{-3/2} \tag{9}$$

I is the ionic strength and z_i is the charge of the species C_i. At $T = 298$ K and for water (dielectric constant $\epsilon_w = 78.5$), A 0.5. Davies has proposed $b = 0.3$ or 0.2 (Stumm and Morgan, 1995). In our calculation, we chose $b = 0.24$ (Morel, 1983).

Comparison of Methods

Newton-Raphson method

Equation 5 is solved with the Newton-Raphson method (Westall, 1979), at the nth iteration with the Jacobian matrix Z^n of the objective functions

$$Z^n_{j,k}\Big|_{\substack{j=1,N_X+N_{CP} \\ k=1,N_X}} = -\frac{\partial Y_j^n}{\partial [X_k]^n}$$

$$Z^n_{j,k}\Big|_{\substack{j=1,N_X+N_{CP} \\ k=N_X-1,N_X+N_{CP}}} = -\frac{\partial Y_j^n}{\partial [CP_{k-N_X}]^n} \tag{10}$$

Z^n can be calculated by two ways. (a) From an analytical computation, we obtain the $(N_X + N_{CP}) * (N_X + N_{CP})$ values of Z^n by

$$Z^n_{j,k}\Big|_{\substack{j=1,N_X \\ k=1,N_X}} = -\sum_{i=1}^{N_C} a_{i,j} \cdot a_{i,k} \frac{[C_i]^n}{[X_k]^n}$$

$$Z^n_{j,k}\Big|_{\substack{j=1,N_X \\ k=N_X+1,N_X+N_{CP}}} = -a\rho_{k-N_X,j}$$

$$Z^n_{j,k}\Big|_{\substack{j=N_X+1,N_X+N_{CP} \\ k=1,N_X}} = a\rho_{k-N_X,j} \cdot \frac{SI_{j-N_X}^n}{[X_k]^n}$$

$$Z^n_{j,k}\Big|_{\substack{j=N_X+1,N_X+N_{CP} \\ k=N_X+1,N_X+N_{CP}}} = 0 \tag{11}$$

Activity coefficients are assumed to be constant for the Jacobian calculation. (b) From the first-order approximation of Eq. 10, we obtain the relationship 12. The progress step of the method ΔX^n (Eq. 13) is obtained by assuming that the objective function Y^{n+1} in Eq. 13 has to equal zero at the $(n+1)^{\text{th}}$ iteration

$$Z^n = \frac{Y^{n+1} - Y^n}{\Delta X^n} \tag{12}$$

$$\Delta X^n = (Z^n)^{-1} \cdot Y^n \tag{13}$$

The values of the component concentrations at the $(n+1)^{\text{th}}$ iteration are

$$[X]^{n+1} = [X]^n + \Delta X^n \tag{14}$$

The new component concentrations are used for calculating the activity coefficients, which will be assumed constant during all the $(n+1)^{\text{th}}$ iteration. Activity correction is then cal-

Table 1. Morel's Tableau for the Gallic Acid Test with pH Fixed to 5.8*

Species	H^+	Al^{3+}	H_3L	log K	Equilib. (M)
H^+	1	0	0	0	1.58×10^{-6}
Al^{3+}	0	1	0	0	2.03×10^{-5}
H_3L	0	0	1	0	2.59×10^{-7}
OH^-	1	0	0	14	6.31×10^{-9}
H_2L^-	−1	0	1	−4.15	1.16×10^{-5}
HL^{2-}	−2	0	1	−12.59	2.65×10^{-8}
L^{3-}	−3	0	1	−23.67	1.39×10^{-13}
$AlHL^+$	−2	1	1	−4.93	2.45×10^{-5}
AlL	3	1	1	9.43	4.90×10^{-4}
AlL_2^{3-}	−6	1	2	−21.98	8.97×10^{-6}
AlL_2^{6-}	−9	1	3	−37.69	1.14×10^{-10}
$Al_2(OH)_2(HL)_3^{3-}$	−8	2	3	−22.65	4.01×10^{-6}
$Al_2(OH)_2(HL)_2L^{3-}$	−9	2	3	−27.81	1.75×10^{-5}
$Al_2(OH)_2(HL)L_2^{4-}$	−10	2	3	−32.87	9.61×10^{-5}
$Al_2(OH)_2L_3^{5-}$	−11	2	3	−39.56	1.24×10^{-4}
$Al_4L_3^{3+}$	−9	4	3	−20.25	2.61×10^{-7}
$Al_3(OH)_4(H_3L)^{4+}$	5	3	1	12.52	6.51×10^{-5}
Total (M)	pH 5.8	10^{-2}	10^{-3}		
Initial (M)	1.58×10^{-6}	variable	variable		
Equilib. (M)	1.58×10^{-6}	2.03×10^{-5}	2.59×10^{-7}		

*Initial concentrations for Al^{3+} and H_3L component are variable.
†Thermodynamic values are from Brassard and Bégurtha (2000).

culated through Picard iterations using concentration values of the n^{th} iteration.

The iteration procedure is used until the convergence criteria for the mass balance (Eq. 15) and for the saturation index when precipitation occurs (Eq. 16) are reached for every component with $\epsilon = 10^{-9}$.

$$\left| \frac{-[T_j] + \sum_{i=1}^{N_C} a_{i,j} \cdot [C_i] + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i]}{[T_j] + \sum_{i=1}^{N_C} |a_{i,j}| \cdot [C_i] + \sum_{i=1}^{N_{CP}} |ap_{i,j}| \cdot [Cp_i]} \right| \leq \epsilon \text{ for } j = 1 \text{ to } N_X \quad (15)$$

$$\left| 1 - Kp_i \prod_{k=1}^{N_{CP}} (\gamma_k \cdot [X_k])^{ap_{i,k}} \right| \leq \epsilon \text{ for } i = 1 \text{ to } N_{CP} \quad (16)$$

The criterion given by Eq. 15 ensures that the solution is obtained with the same precision for every component in the system, whatever are the low or very high concentrations.

A singular or near-singular Jacobian matrix of the objective functions (Reed, 1982) is unfortunately not the only condition for the nonconvergence of the Newton-Raphson method. Walsh et al. (1984) have reported that poor initial root guesses may cause convergence difficulties, and Brassard and Bodurtha (2000) have recently shown the high sensitivity of first- and second-order methods with respect to the initial vector of components for gallic acid (Table 1). For many starting points (Figure 1a), the Newton-Raphson method is blocked in a loop and turns infinitely (Figure 2a).

Nonconvergence of the Newton-Raphson method can occur in a field of great environmental importance. Pyrite (FeS_2), which is the most abundant metal sulfide on Earth, is involved in many processes such as formation of acid mine drainage, redox cycling of metals in sediments, oxic-anoxic boundaries of sediments or estuaries, flotation, and nuclear waste disposals (Singer and Stumm, 1970; Morse et al., 1987;

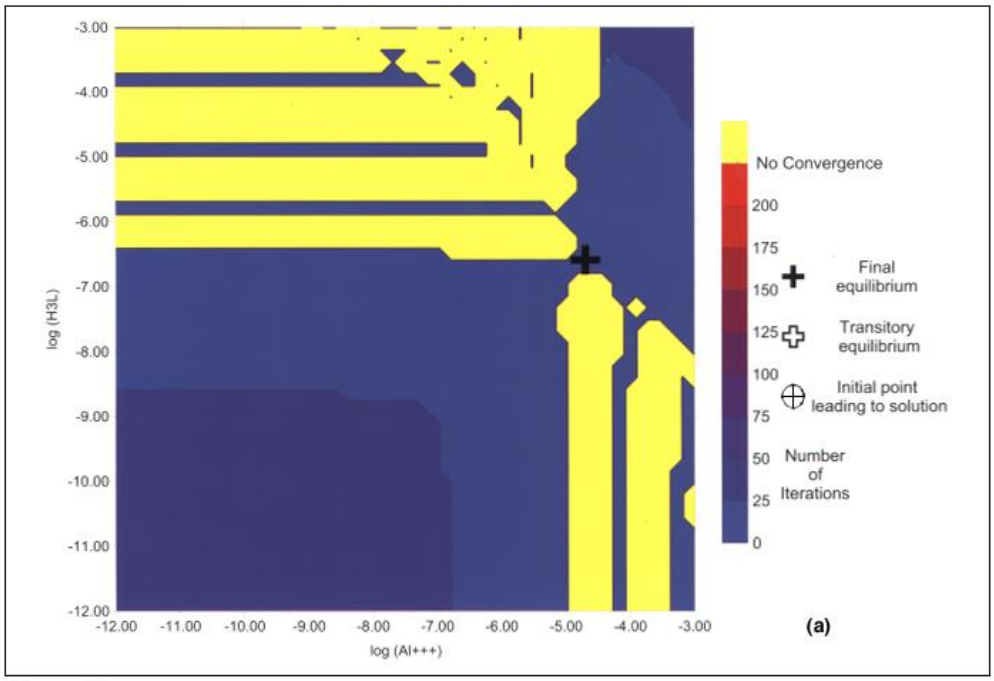


Figure 1. Convergence map.

They report if the solution is achieved and provide the number of iterations needed for convergence. X- and Y-axes correspond to the 50×50 initial component concentration vectors used to initiate the system's resolution. (a) Resolution of the gallic acid test with the Newton-Raphson method starting in nonconvergence zones (yellow) trapped in the infinite loop; (b) resolution of the FeS_2 test with the Newton-Raphson method (with initial concentrations of 0.1 μm for H^+ and 1 mM for O_2). Seven starting points only lead to the solution after around 250 iterations. Transitory equilibrium is calculated first. Final equilibrium is reached for the precipitation of FeS_2; (c) resolution of the FeS_2 test with the Newton-Raphson method on CAI (with initial concentrations of 0.1 μM for H^+ and 1 mM for O_2). Only some starting points induce nonconvergence; (d) resolution of the gallic acid test with the combined algorithm with convergence reaching starting points. Blue zones (before yellow) show the influence of the preconditioning method, red ones show the interest of the reconditioning method. (Figure continued on next page).

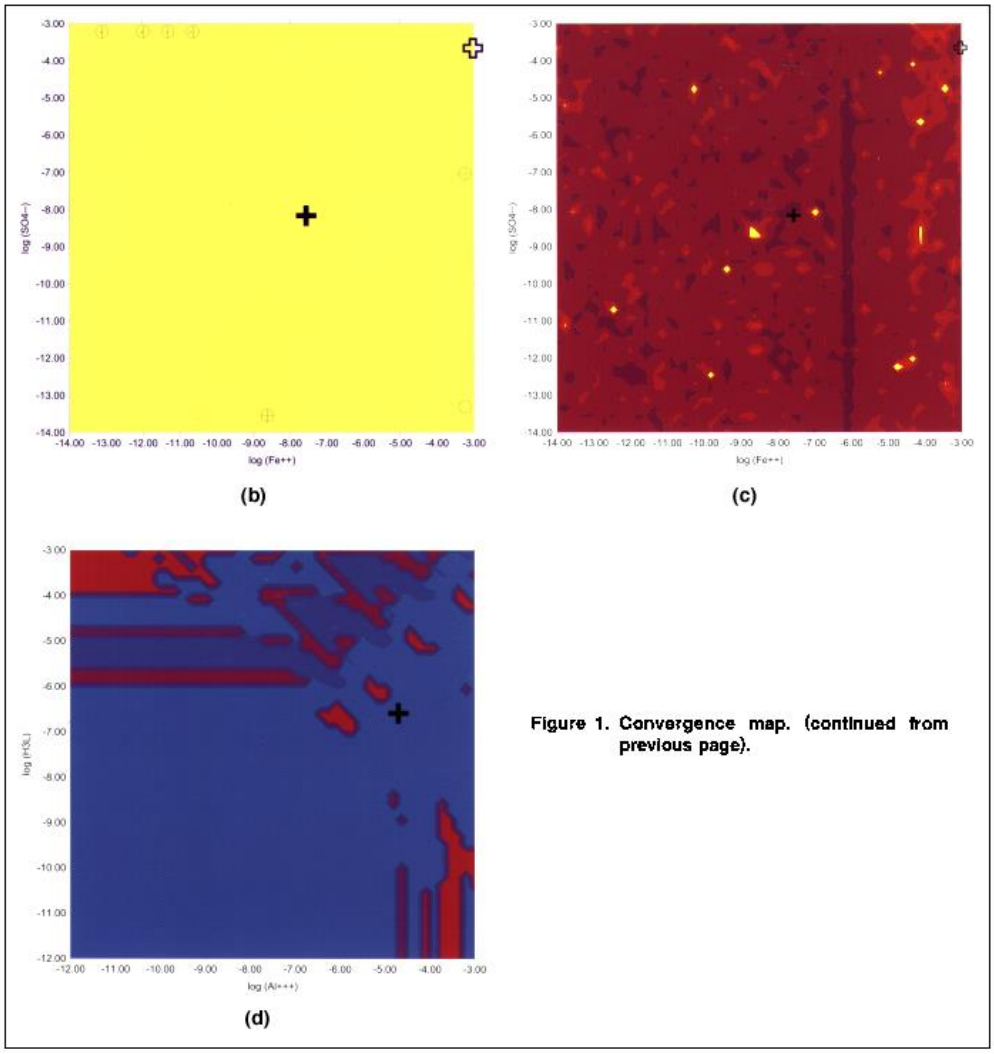


Figure 1. Convergence map. (continued from previous page).

Chilakapati, 1999; Lalvani et al., 1991; Bosch, 1999). It is therefore very important to calculate the FeS_2 oxidation without any convergence problem knowing the complexity of thermodynamic equilibria. The use of analytical solutions avoids a convergence problem but leads to severe simplifications of a chemical problem, as shown by comparing the three major reactions analytically used by Chilakapati (1999) and the complexity of the 52 species reported in Table 2. The system is nonideal and activity are corrected using Davies approximation (Eq. 7). Nevertheless, complexity induces convergence problems. The inefficiency of the Newton-Raphson method is mainly due to the initial dissolved oxygen concentration (Figure 1b). In such a case, the system is assumed to be closed without any contact with the atmosphere. The infiltrated meteoritic water is assumed to be in equilibrium with the atmospheric oxygen, that is, initial O_2 concentration = 1 mM. The total concentration of the component O_2 is negative to respect the mass balance for the formation of 1 mmol

FeS_2 in one liter with the chosen set of components

$$Fe^{2+} + 2SO_4^{2-} + 2H^+ - \frac{7}{2}O_2 \rightarrow FeS_2 + H_2O \quad (17)$$

The component H_2O, which is the solvent, was omitted in Table 2, because its activity is constant and equal to one. As shown in Figure 2b, the Newton-Raphson method can give excessive concentrations, which are physically unacceptable.

Chemically allowed interval (CAI)

It has been proven (Weltin, 1990) that the objective function (Eq. 5) has a unique solution in the CAI, defined as $0 < [X_j] < T_j$. However, this definition is restricted to chemical systems where all stoichiometric coefficients are positive. For the most general chemical system, we define the CAI as $0 < [X_j] < Max_j$ and set a definition of the upper limit (Max_j) of the CAI based on the limiting reactive notion

$$\mathrm{Max}_j = \left[T_j\right] + \sum_{i=1}^{N_C} \left|a_{i,j}\right|_{a_{i,j}<0} \cdot \left(\min\left|a_{i,k} \cdot [T_k]\right|_{a_{i,k}>0}\right) \quad (18)$$

It can be seen in the *FeS*$_2$ test-case (Table 2) that the Newton-Raphson method can overstep the CAI and then can give

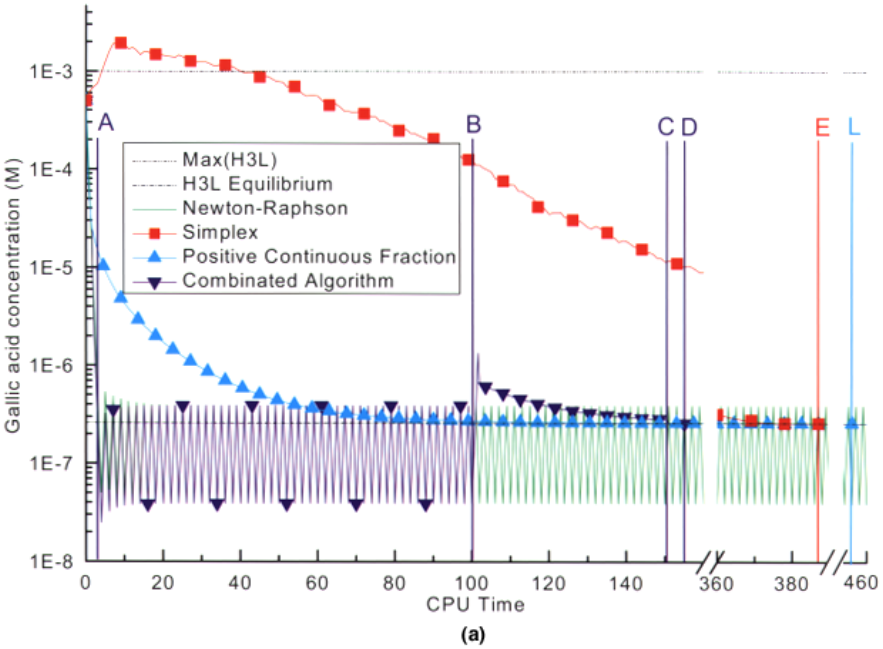


Figure 2. Newton-Raphson, Newton-Raphson on CAI, simplex, positive continuous fraction methods, and the combined algorithm.

X-axis gives the CPU time; unity is the time needed for one Newton-Raphson iteration. Line A corresponds to iteration when the combined algorithm leaves preconditioning with positive continuous fraction method for searching the solution with the Newton-Raphson method. Line D represents the end point of the combined algorithm with the Newton-Raphson method: (a) gallic acid test (from pH 5.8, 0.01 nM Al^{3+} and 0.5 mM H_3L). Line A, iteration 2; Line B, iteration 101, where the combined algorithm overruns maximum iterations allowed for the Newton-Raphson, and reconditioning by positive continuous fraction method; Line C, iteration 151, where the combined algorithm is close enough to the solution and leaves reconditioning by positive continuous fraction method for searching the solution with the Newton-Raphson method; Line D, iteration 155; Line E, iteration 392 corresponds to convergence with the Simplex method; Line L, iteration 457; convergence of the positive continuous fraction method. The Newton-Raphson method does not converge. (b) FeS_2 test (from pH 7, 1 mM O_2, 1 mM Fe^{2+} and 2 mM SO_4^{2-}). Line A, iteration 7; Line D, iteration 48; Line F, iteration 16 corresponds to the convergence to the transitory equilibrium of the combined algorithm with the Newton-Raphson method. Line G, iteration 193, where convergence to the transitory equilibrium of the Newton-Raphson method. Line H, iteration 225, where the Newton-Raphson method finally converges; Line I, iteration 108, where the Newton-Raphson method with respect of the CAI converges to the transitory equilibrium; Line J, iteration 140, where the Newton-Raphson method with respect of the CAI finally converges; Line K, iteration 300, where convergence to the transitory equilibrium of the positive continuous fraction method; iteration of 2 millions corresponds to convergence of the positive continuous fraction method. The Newton-Raphson method does not converge. The Simplex method does not converge and then stays blocked at the point: 0.132 μM H^+, 1.2 mM O^2, 1.22 mM Fe^{2+} and 0.995 mM SO_4^{2-}. Convergence with the Newton Raphson method is obtained if the starting point is: pH 7, 1 mM O_2, 596.4 μM Fe^{2+}, and 91.03 nM SO_4^{2-}.

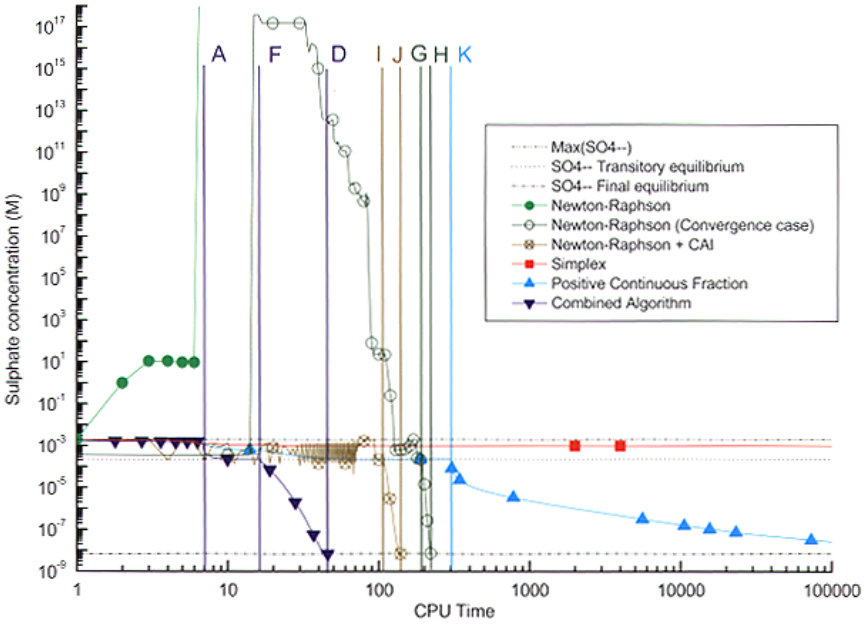


Figure 2. Newton-Raphson, Newton-Raphson on CAI, Simplex, positive continuous fraction methods, and the combined algorithm. (Continued from previous page).

concentrations which are much too high (see Figure 2b). Imposing a CAI appears thus as an important condition in a robust algorithm. The lowest limit of this interval, zero, is intuitive and has been taken into account by many computer codes. Morel and Morgan (1972) have proposed the following rule which has been used by Westall (1979) for the computer code MICROQL

$$\text{if } \left[X_j\right]^n + \Delta X_j^n < 0 \text{ then } \left[X_j\right]^{n+1} = \frac{\left[X_j\right]^n}{10} \qquad (19)$$

We propose the new rule to impose the upper limit of the CAI

$$\text{if } \left(\left[X_j\right]^n + \Delta X_j^n > \text{Max}_j\right) \text{ then } \left(\log\left(\left[X_j\right]^{n+1}\right) = \frac{\log\left(\text{max}_j\right) + \log\left(\left[X_j\right]^n\right)}{2}\right) \qquad (20)$$

For the H^+component, to impose the pH in the usual domain, the lowest limit of the CAI is 10^{-14} and the upper limit is the smallest value 1 or Max_{H^+} as computed from Eq. 18. From these definitions, the CAI for the FeS_2 test case is between 10^{-14} and 32.5 mM, 0 and 286.3 mM, 0 and 1 mM, and 0 and 2 mM for components H^+, O_2, Fe^{2+}, and SO_4^{2-}, respectively.

By imposing the CAI in the Newton-Raphson algorithm, we avoid the risk of major overestimation in the FeS_2 test case (Figure 2b). The solution is then reached for a majority of starting points (see Figure 1c). This improvement is easy to introduce and very inexpensive in terms of computing time. Although very efficient for some cases, no improvement of convergence can be observed if the loop phenomena belongs to the CAI as, for example, the gallic acid test (Figure 2a).

Continuous fraction method

The continuous fraction method has been used to solve thermodynamic equilibrium in the computer code WATSPEC (Wigley, 1977), or for pre-conditioning the Newton-Raphson method for the major species in the PHREEQC code (Parkhurst and Appelo, 1999). This method, which only needs one computation of the approximate thermodynamic equilibrium per iteration, is the cheapest zero-order method. The iteration procedure is

Table 2. Morel's Tableau for the FeS_2 Test*

Species	H^+	O_2	Fe^{2+}	SO_4^{2-}	log K	Equilibrium (M)	Activity Coefficient
Aqueous							
H^+	1	0	0	0	0	9.24×10^{-8}	0.9998
OH^-	−1	0	0	0	−14	1.08×10^{-7}	0.9998
O_2	0	1	0	0	0	0.0	1
Fe^{2+}	0	0	1	0	0	2.73×10^{-8}	0.9991
$Fe(OH)_2$	−2	0	1	0	−20.6	8.03×10^{-15}	1
$Fe(OH)_3^-$	−3	0	1	0	−31	3.46×10^{-8}	0.9998
$FeOH^+$	−1	0	1	0	−9.5	9.35×10^{-11}	0.9998
$FeSO_4$	0	0	1	1	2.2	2.96×10^{-14}	1
Fe^{3+}	1	0.25	1	0	8.49	0.0	0.9980
$Fe(OH)_2^+$	−1	0.25	1	0	2.82	1.12×10^{-16}	0.9998
$Fe(OH)_3$	2	0.25	1	0	3.5	5.68×10^{16}	1
$Fe(OH)_4^-$	−3	0.25	1	0	−13.11	1.55×10^{-8}	0.9998
$FeOH^{2+}$	0	0.25	1	0	6.3	3.14×10^{-20}	0.9991
$Fe_2(OH)_2^{4+}$	0	0.5	2	0	14.03	0.0	0.9965
$Fe_3(OH)_4^{5+}$	−1	0.75	3	0	19.17	0.0	0.9946
$Fe(SO_4)_2^-$	1	0.25	1	2	11.7	0.0	0.9998
$FeSO_4^{+}$	1	0.25	1	1	10.4	0.0	0.9998
SO_4^{2-}	0	0	0	1	0	6.86×10^{-9}	0.9991
HSO_4^-	1	0	0	1	1.98	6.05×10^{-14}	0.9998
H_2SO_4	2	0	0	1	−1.02	0.0	1
SO_3^{2-}	0	−0.5	0	1	−46.62	4.97×10^{-19}	0.9991
HSO_3^-	1	−0.5	0	1	−39.42	7.27×10^{-19}	0.9998
H_2SO_3	2	−0.5	0	1	−37.41	0.0	1
SO_2	2	0.5	0	1	37.56	0.0	1
$HS_2O_3^-$	3	−2	0	2	−132.52	0.0	0.9998
$S_2O_3^{2-}$	2	−2	0	2	−133.54	9.64×10^{-19}	0.9991
H_2S	2	2	0	1	131.33	2.28×10^{-8}	1
HS^-	1	−2	0	1	−138.32	2.52×10^{-8}	0.9998
S^{2-}	0	−2	0	1	−151.25	3.21×10^{-24}	0.9991
S_2^{2-}	2	−3.5	0	2	−243.37	3.93×10^{-19}	0.9991
S_3^{2-}	4	−5	0	3	−335.56	0.0	0.9991
S_4^{2-}	6	−6.5	0	4	−427.97	0.0	0.9991
S_5^{2-}	8	8	0	5	520.60	0.0	0.9991
$S_2O_4^{2-}$	2	−1.5	0	2	−118.46	0.0	0.9991
$S_2O_5^{2-}$	2	1	0	2	−83.65	0.0	0.9991
$S_2O_7^{2-}$	2	0.5	0	2	51.42	0.0	0.9991
$S_2O_8^{2-}$	2	0.5	0	2	−22.5	0.0	0.9991
$S_3O_8^{2-}$	4	−2	0	3	−146.1	0.0	0.9991
$S_4O_6^{2-}$	6	−0.35	0	4	−22.88	0.0	0.9991
$S_5O_6^{2-}$	8	−5	0	5	−332.54	0.0	0.9991
Mineral							
Fe(s)	−2	−0.5	1	0	−59.03	0.0	
S(s)	2	1.5	0	1	93.22	0.0	
$Fe(OH)_2$	−2	0	1	0	−13.90	0.0	
$Fe(OH)_3$	−2	0.25	1	0	2.83	0.0	
$Fe(SO_4)_3$	2	0.5	2	3	13.77	0.0	
FeO	−2	0	1	0	−13.53	0.0	
$FeSO_4$	0	0	1	2	−2.66	0.0	
Goethite	−2	0.25	1	0	7.95	0.0	
Hematite	−4	0.5	2	0	16.87	0.0	
Melanterite	0	0	1	1	2.35	0.0	
Pyrite	2	−3.5	1		−217.40	9.9997×10^{-4}	
Pyrrhotite	0	−2	1	1	−134.6	0.0	
Wustite	−1.894	0.0265	0.947	0	−11.51	0.0	
Total (M)	2×10^{-3}	-3.5×10^{-3}	1×10^{-3}	2×10^{-3}			
Initial (M)	1×10^{-7}	1×10^{-3}	variable	variable			
Transitory equilib. (M)	1.93×10^{-8}	3.35×10^{-74}	9.58×10^{-4}	2.23×10^{-4}			
Equilib. (M)	9.24×10^{-8}	1.10×10^{-73}	2.73×10^{-8}	6.86×10^{-9}			

* Total concentrations correspond to 1 mmol FeS_2 dissolution in 1 L pure water. Equilibrium is obtained after precipitation of pyrite. Activity is corrected using Davies approximation. Thermodynamic values are from CHESS data base (van der Lee, 1998); Ionic strength $I = 1.81\times10^{-7}$ at $T = 298$ K. Activity of pure mineral = 1.

$$[X_j]^{n+1} = [X_j]^n \cdot \frac{[T_j]}{\sum_{i=1}^{Nc} a_{i,j} \cdot [C_i]^n} \quad (21)$$

However, the method is inoperative for systems in which a component has a zero total concentration $[T_j]$, as can be seen in relationship 21. The method can also diverge if some stoichiometric coefficients are negative. Often, the component H^+ has a zero total concentration and is associated with negative stoichiometric coefficients. In the code WATSPEC (Wigley, 1977), the pH value must be imposed to find the thermodynamic equilibrium. Hydrogen and oxygen are excluded from the continuous fraction pre-conditioning in the code PHREEQC (Parkhurst and Appelo, 1999). Moreover, it has never been used for nonideal system.

Positive continuous fraction method

We define two new values to take into account a component with zero (H^+ case) or negative (ion-exchange) total concentration, and to be more efficient with negative stoichiometric coefficients. The *reactive sum* is defined by Eq. 22 or 24 and the *product sum* by Eq. 23 or 25. The *reactive* (respectively, *product*) adjective refers to the C_i species where the component X_j is a reactive, that is, $a_{i,j} > 0$ (respectively, product, that is, $a_{i,j} < 0$).

If $[T_j] \geq 0$

$$\mathrm{Sum}_j^{\mathrm{reac}} = \sum_{a_{i,j}>0} a_{i,j} \cdot [C_i] \quad (22)$$

$$\mathrm{Sum}_j^{\mathrm{prod}} = [T_j] + \sum_{a_{i,j}<0} |a_{i,j}| \cdot [C_i] \quad (23)$$

If $[T_j] < 0$

$$\mathrm{Sum}_j^{\mathrm{reac}} = |[T_j]| + \sum_{a_{i,j}>0} a_{i,j} \cdot [C_i] \quad (24)$$

$$\mathrm{Sum}_j^{\mathrm{prod}} = \sum_{a_{i,j}<0} |a_{i,j}| \cdot [C_i] \quad (25)$$

Using these two new values, the mass balance Eq. 3 is written, without precipitation, at equilibrium

$$\mathrm{Sum}_j^{\mathrm{reac}} = \mathrm{Sum}_j^{\mathrm{prod}} \quad (26)$$

We choose a species C_{i0}, for which the stoichiometric coefficient $a_{i0,j}$ for component X_j is not zero. Very often, C_{i0} can be chosen equal to X_j. For other cases, we propose to take $a_{i0,j}$ as the smallest value of the strictly positive stoichiometric coefficient. The mass action laws are written for the reactive sum if $a_{i0,j}$ is positive (respectively, for the product sum if $a_{i0,j}$ is negative) by using component concentrations at iterations n and $(n+1)$

$$\left(\gamma_j [X_j]^{n+1}\right)^{a_{i0,j}} \cdot \left[\sum_{a_{i,j}>0} a_{i,j} \cdot K_i \prod_{k \neq j} \left(\gamma_k [X_k]^n\right)^{a_{i,k}} \cdot \left(\gamma_j [X_j]^n\right)^{a_{i,j} - a_{i0,j}} \right] = [T_j] + \sum_{a_{i,j}<0} |a_{i,j}| \cdot [C_i]^n \quad (27)$$

After reordering, it becomes

$$\left(\gamma_j [X_j]^{n+1}\right)^{a_{i0,j}} = \frac{\left(\gamma_j [X_j]^n\right)^{a_{i0,j}}}{\left(\gamma_j [X_j]^n\right)^{a_{i0,j}}} \times \frac{[T_j] + \sum_{a_{i,j}<0} |a_{i,j}| \cdot [C_i]^n}{\sum_{a_{i,j}>0} a_{i,j} \cdot K_i \prod_{k \neq j} \left(\gamma_k [X_k]^n\right)^{a_{i,k}} \cdot \left(\gamma_j [X_j]^n\right)^{a_{i,j} - a_{i0,j}}} \quad (28)$$

Since the product and reactive sums appear in Eq. 28, we obtain the relationship 29 giving $[X_j]$ at the $(n+1)^{\mathrm{th}}$ iteration

$$[X_j]^{n+1} = [X_j]^n \cdot \left(\frac{\mathrm{Sum}_j^{\mathrm{prod},n}}{\mathrm{Sum}_j^{\mathrm{reac},n}} \right)^{1/a_{i0,j}} \quad (29)$$

In relationship 29, the impact of the modification of the concentration $[X_k]$ over the Y_j function, as with all the zero-order methods, is not taken into account. The simultaneous modification of all component concentrations can induce unfavorable oscillations. In the positive continuous fraction method, the weighted mean given in relationship 30 is then used to calculate the Nx component concentrations at the $(n+1)^{\mathrm{th}}$ iteration, with θ between 0 and 1. This ensures convergence for certain

$$[X_j]^{n+1} = \theta \cdot [X_j]^n \cdot \left(\frac{\mathrm{Sum}_j^{\mathrm{prod},n}}{\mathrm{Sum}_j^{\mathrm{reac},n}} \right)^{1/a_{i0,j}} + (1 - \theta) \cdot [X_j]^n \quad (30)$$

For a high θ value, such as $\theta = 0.9$, the near solution is reached quickly. Indeed, this corresponds to dividing (respectively, multiplying) the $[X_j]$ value per 10 if this value is much higher (respectively, lower) than the equilibrium value. Close to the solution, a small θ value ($\theta = 0.1$) avoids oscillations and ensures convergence. θ is thus calculated as an adaptive parameter

$$\text{if } \left(\mathrm{Sum}_j^{\mathrm{reac}} > \mathrm{Sum}_j^{\mathrm{prod}}\right) \text{ then } \left(\theta_j = 0.9 - \frac{\mathrm{Sum}_j^{\mathrm{prod}}}{\mathrm{Sum}_j^{\mathrm{reac}}} \cdot 0.8\right)$$

$$\text{if } \left(\mathrm{Sum}_j^{\mathrm{reac}} < \mathrm{Sum}_j^{\mathrm{prod}}\right) \text{ then } \left(\theta_j = 0.9 - \frac{\mathrm{Sum}_j^{\mathrm{reac}}}{Sum_j^{\mathrm{prod}}} \cdot 0.8\right) \quad (31)$$

If there is some precipitated species, the procedure is the same as presented in the modeling part, that is, research of the transitory equilibrium without precipitation, calculation of the SI, precipitation of the more supersaturated species Cp_i, and research of a new equilibrium.

We choose a component X_j which is reactive ($ap_{i,j}$ positive) for the precipitation of Cp_i. The precipitated species Cp_i controls the mass balance relationship for the component X_j

$$[Cp_i]^{n+1} = [T_j] - \frac{1}{a_{i,j}} \left(\sum_{k=1}^{N_C} a_{k,j} [C_k]^n - \sum_{\substack{k=1 \\ k \neq i}}^{N_{CP}} ap_{k,j} [Cp_k]^n \right) \tag{32}$$

The component X_j controls the saturation index of species Cp_i through the new definition of the *product sum*

$$\mathrm{Sum}_j^{\mathrm{prod}} = 1 \tag{33}$$

and the *reactive sum*

$$\mathrm{Sum}_j^{\mathrm{reac}} = \left(Kp_i \prod_{k=1}^{N_x} \left(\gamma_k [X_k]^n \right)^{ap_{i,k}} \right)^{1/ap_{i,j}} \tag{34}$$

Equations 22 to 25 are used for the other components. The iteration procedure to find the current equilibrium is first Eq. 32 for all the precipitated species, and then Eq. 30 for all the components with θ calculated with Eq. 31. Convergence is obtained if

$$\frac{\mathrm{Sum}_j^{\mathrm{reac}} - \mathrm{Sum}_j^{\mathrm{prod}}|}{\mathrm{Sum}_j^{\mathrm{reac}} + \mathrm{Sum}_j^{\mathrm{prod}}} \leq \epsilon \tag{35}$$

for all the components.

Since the positive continuous fraction method does not need computation and inversion of the Jacobian of objective functions, one iteration is much faster with this new method than with the Newton-Raphson or other first-order methods, such as twice for the gallic acid test and 10 times faster for the FeS_2 test. Moreover, the positive continuous fraction method very quickly reaches close solution, but it takes some time to get a precise convergence (Figure 2a), whereas the Simplex method takes a long time far from the solution before finding the right way and quickly converging. This method appears to be very attractive for approximating the solution. Moreover, the solution of the FeS_2 test can be obtained with the positive continuous fraction method when other methods cannot (Figure 2b). The Simplex or Newton-Raphson methods are based on the slope of the objective function to search the solution. For initial value of the O_2 component concentration close to 1 mM, both methods cannot converge due to high iron and sulfate values. On the contrary, the positive continuous fraction method, which does not depend on the slope of the objective function, is thus insensitive to local minima and infinite loop phenomena (Figure 2) and gives the solution of the gallic and the FeS_2 tests with whatever initial point is used.

A New Efficient Combined Algorithm

The positive continuous fraction method is a new, very robust and fast zero-order numerical method to approach thermodynamic equilibrium of complex chemical systems. By coupling this method with a first-order method which quickly converges near the solution, a combined algorithm, impressive with respect to the criteria of reliability, robustness, and quickness, is then obtained. In order to be close to the solution and use the Newton-Raphson method in the best conditions, a first step to the solution with the positive continuous fractions method (pre-conditioning) is performed under the assumption of an ideal system. When the error is less than 50%, the accurate solution is searched with the Newton-Raphson method (see Figure 2b) associated with the respect of the CAI for the real system. Some nonconvergence areas can be very close to the solution (Figure 1a), and the pre-conditioning is thus inefficient, as shown in Figures 1d and 2a. If the Newton-Raphson method does not converge after the maximum allowed number of iterations, the solution research continues with the positive continuous fraction method (re-conditioning) for the real system. For a relative error less than 5%, the final solution is then found with the fast Newton-Raphson method. For the FeS_2 test during the pre-conditioning, the O_2 value, which is responsible for the nonconvergence, quickly decreases. In this case, the solution for every initial values reported in Figure 1b is found with the combined algorithm. Moreover, the CPU time is five times faster (48 iterations) than with the Newton-Raphson method (250 iterations) for the seven initial values where convergence is observed (Figure 2b). The efficiency of the re-conditioning is shown from the gallic acid test for which the solution is found for every starting point (Figure 1d).

This new algorithm, which makes the resolution of highly nonlinear algebraic chemical system more robust, rapid, and without any constraint about the initial values, should also be useful for the resolution of all nonlinear algebraic systems.

Conclusion

In this article, we thus propose to associate a new method, the positive continuous fraction, a new constraint, respecting the chemically allowed interval (CAI) and the Newton-Raphson method, for solving nonlinear algebraic systems which cannot be solved using the classical Newton-Raphson method. Efficiency of the positive continuous fractions method for pre- or re-conditioning and of the CAI for robustness are shown in their association with the simplest Newton-Raphson method in the combined algorithm. For each specific problem, an adapted relaxation technique can be implemented in the Newton-Raphson method. The new combined algorithm will thus be strengthened and accelerated by the use of the efficient Positive Continuous Fraction pre- or re-conditioning, by imposing the CAI and by the specific relaxation technique.

The implementation of this combined algorithm in reactive transport code instead of the classical ones shall highly increase the robustness of the chemical module. This algorithm converges even if initial conditions are far from equilibrium. This allows greater transport time steps and the modeling of very sharp concentration fronts. Some subsequent computing time reduction shall be induced: each thermodynamic equilibrium is computed faster, and the increase in the transport time steps should reduce their number. Finally, the reactive transport code which includes this combined algorithm shall be able to induce new predictions which are more detailed

and, over a longer period, in sensible domains such as contaminant migration or nuclear waste disposal assessment.

Acknowledgments

We thank referees for their helpful comments. This work is supported by the Programme Environnement, Vie & Sociétés of CNRS and a grant from Ministère de l'Education Nationale, de la Recherche et de la Technologie to J. C.

Notation

$a_{i,j}$ = stoichiometric coefficient of component X_j for the formation of species C_i
$ap_{i,j}$ = stoichiometric coefficient of component X_j for the formation of precipitated species Cp_i
A = parameter for Davies activity correction ($A \approx 0.5$ at 298 K)
b = parameter for Davies activity correction ($b = 0.24$ at 298 K)
C_i = species
Cp_i = precipitated species
$[Cp_i]$ = amount of precipitated species Cp_i per liquid volume unit, M
$[C^\circ]$ = reference concentration $[C^\circ] = 1$ M
I = ionic strength
K_i = equilibrium constant for the formation of species C_i
Kp_i = precipitation constant for the formation of precipitated species Cp_i
Ks_i = solubility product of precipitated species Cp_i
Max_j = upper limit of the chemically allowed interval for component X_j, M
N_c = number of species
N_{CP} = number of precipitated species
N_X = number of components
SI_i = saturation index of precipitated species Cp_i
$\mathrm{Sum}_j^{\mathrm{prod}}$ = product sum for component X_j, M
$\mathrm{Sum}_j^{\mathrm{reac}}$ = reactive sum for component X_j, M
T = temperature ($T = 298$ K)
$[T_j]$ = total concentration of component X_j, M
X_j = component
Y_j = objective function for component X_j, M
$T_{j=N_X+i}$ = objective function for precipitated species Cp_i
z_i = electric charge of species C_i
Z^n = Jacobian matrix of the objective functions at iteration n
$Z^n_{j,k}$ = Jacobian matrix coefficient at iteration n
γ_i = activity coefficient of species C_i
ΔX^n = progress step of the Newton-Raphson method at iteration n (M)
ϵ = precision criterion of the methods ($\epsilon = 10^{-9}$)
ϵ_w = dielectric constant of water ($\epsilon_w = 78.5$)
ϑ = weighted mean coefficient for the Positive Continuous Fraction method
i = subscript of species
j,k = subscript of component
n = superscript of iteration
[ ] = concentration, M
{ } = activity

Literature Cited

Bellot, J. C., R. V. Tarantino, and J. S. Condoret, "Thermodynamic Modeling of Multicomponent Ion-Exchange Equilibria of Amino Acids," *AIChE J.*, **45**, 1326 (1999).

Bosch, X., "Doñana Clean-up 'Left Half the Soil Still Contaminated'," *Nature*, **398**, 178 (1999).

Brassard, P., and P. Bodurtha, "A Feasible Set for Chemical Speciation Problems," *Comput. Geosci.*, **26**, 277 (2000).

Bryant, S. L., R. S. Schechter, and L. W. Lake, "Interactions of Precipitation/Dissolution Waves and Ion Exchange in Flow through Permeable Media," *AIChE J.*, **32**, 751 (1986).

Chilakapati, A., "Optimal Design of a Subsurface Redox Barrier," *AIChE J.*, **45**, 1342 (1999).

Kersting, A. B., D. W. Efud, D. L. Finnegan, D. J. Rokop, D. K. Smith, and J. L. Thompson, "Migration of Plutonium in Ground Water at the Nevada Test Site," *Nature*, **397**, 56 (1999).

Krebs, R., M. Sardin, and D. Schweich, "Mineral Dissolution, Precipitation and Ion Exchange in Surfactant Flooding," *AIChE J.*, **33**, 1371 (1987).

Lalvani, S. B., B. A. DeNeve, and A. Weston, "Prevention of Pyrite Dissolution in Acidic Media," *Corrosion*, **47**, 55 (1991).

Morel, F. M. M., and J. J. Morgan, "A Numerical Method for Computing Equilibria in Aqueous Chemical Systems," *Environ. Sci. Technol.*, **6**, 58 (1972).

Morel, F. M. M., *Principles of Aquatic Chemistry*, Wiley Interscience, New York, 446 pp. (1983).

Morin, K. A., "Simplified Explanations and Examples of Computerized Methods for Calculating Chemical Equilibrium in Water," *Comput. Geosci.*, **11**, 409 (1985).

Morse, J. W., F. J. Millero, J. C. Cornwall, and D. Rickard, "The Chemistry of the Hydrogen Sulfide and Iron Sulfide Systems in Natural Waters," *Earth Sci. Rev.*, **24**, 1 (1987).

Nelder, J. A., and R. Mead, "A Simplex Method for Function Minimization," *Comput. J.*, **7**, 308 (1965).

Nordstrom, D. K., and J. W. Ball, "Chemical Models, Computer Programs and Metal Complexation in Natural Waters," *Complexation of Trace Metals in Natural Waters*, C. J. M. Kramer and J. C. Duinker, eds., M. Nijhoff and W. Junk (Publishers), The Hague, The Netherlands (1984).

Parkhurst, D. L., and C. A. J. Appelo, "User's Guide to PHREEQC (version 2)—A Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations," Water-Resour. Invest. Rep. 99-4259, U.S. Geological Survey, Denver, CO (1999a).

Parkhurst, D. L., and C. A. J. Appelo, "PHREEQC: http://water.usgs.gov/software/phreege.htlm (1999b).

Reed, M. H., "Calculation of Multicomponent Chemical Equilibria and Reaction Processes in Systems Involving Minerals, Gases and an Aqueous Phase," *Geochem. Cosmochem. Acta*, **46**, 513 (1982).

Singer, P. C., and W. Stumm, "Acid Mine Drainage: The Rate Limiting Step," *Sci.*, **167**, 1121 (1970).

Stumm, W., and J. J. Morgan, *Aquatic Chemistry*, 3rd ed., Wiley-Interscience, New York, 1,040 pp. (1995).

Syed, F. H., R. Datta, and K. L. Jensen, "Thermodynamically Consistent Modeling of a Liquid-Phase Nonisothermal Packed-Bed Reactor," *AIChE J.*, **46**, 380 (2000).

van der Lee, J., "Thermodynamic and Mathematical Concepts of CHESS," Technical Report Nr. LHM/RD/98/39, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France (1998a).

van der Lee, J., "CHESS: http://chess.ensmp.fr" (1998b).

Walsh, M. P., S. L. Bryant, R. S. Schechter, and L. W. Lake, "Precipitation and Dissolution of Solids Attending Flow through Porous Media," *AIChE J.*, **30**, 317 (1984).

Weltin, E., "Are the Equilibrium Concentrations for a Chemical Reaction Always Uniquely Determined by the Initial Concentrations?" *J. Chem. Educ.*, **67**, 548 (1990).

Westall, J. C., "MICROQL: a Chemical Equilibrium Program in BASIC. Computation of Adsorption Equilibria in BASIC," Swiss Federal Institute of Technology EAWAG, Dübendorf, Switzerland (1979).

White, A. F., J. M. Delany, T. N. Narasimhan, and A. Smith, "Groundwater Contamination from an Inactive Uranium Mill Tailings Pile: 1. Application of a Chemical Mixing Model," *Water Resour. Res.*, **20**, 1743 (1984).

Wigley, T. M. L., "WATSPEC: A Computer Program for Determining the Equilibrium Speciation of Aqueous Solutions," Brit. Geomorphol. Res. Group Tech. Bull. 20 (1977).

Wood, J. R., "Calculation of Fluid-Mineral Equilibria Using the Simplex Algorithm," *Comput. Geosci.*, **19**, 23 (1993).

Manuscript received June 12, 2001, and revision received Oct. 12, 2001.

Annexe 4. Comparison of linear solvers for equilibrium geochemistry computations

Comput Geosci (2017) 21:131–150
DOI 10.1007/s10596-016-9600-5

ORIGINAL PAPER

Comparison of linear solvers for equilibrium geochemistry computations

Hela Machat[1,2,3] · **Jérôme Carrayrou**[1]

Received: 11 February 2015 / Accepted: 24 October 2016 / Published online: 10 December 2016

Abstract Equilibrium chemistry computations and reactive transport modelling require the intensive use of a linear solver under very specific conditions. The systems to be solved are small or very small (4×4 to 20×20, occasionally larger) and are very ill-conditioned (condition number up to 10^{100}). These specific conditions have never been investigated in terms of the robustness, accuracy, and efficiency of the linear solver. In this work, we present the specificity of the linear system to be solved. Several direct and iterative solvers are compared using a panel of chemical systems, including or excluding the formation of mineral species. We show that direct and iterative solvers can be used for these problems and propose computational keys to improve the chemical solvers.

Keywords Geochemical modelling · Instantaneous equilibrium chemistry · Linear system inversion · Linear solver · Small matrix · Ill-conditioned matrix · Newton-Raphson algorithm

Electronic supplementary material The online version of this article (doi:10.1007/s10596-016-9600-5) contains supplementary material, which is available to authorized users.

✉ Jérôme Carrayrou
jerome.carrayrou@unistra.fr

[1] CNRS, ENGEES, LHyGeS UMR 7517, Université de Strasbourg, 67000 Strasbourg, France

[2] Ecole Supérieure des Ingénieur de l'Equipement Rural de Medjez el Bab, Université de Jendouba, Jendouba, Tunisia

[3] Université de Monastir, UR13ES63-Chimie Appliquée et Environnement, 5000 Monastir, Tunisia

1 Introduction

The problem of groundwater management is receiving increasing attention, and many tools have been developed to address this issue. One of these tools, reactive transport models, was first limited to laboratory experiments and was then extended to field problem comprehension. In recent decades, reactive transport models have increased in complexity and efficiency, and they are now used in many fields. Reactive transport models have been used to study the transport of contaminants, such as heavy metals [1, 2] and radioelements [3–5]. Because of the increasing interest in questions related to climate change, many studies on reactive transport have been conducted to examine the possibility of geologic CO_2 sequestration [6–10].

Under the wide variety of models and cases lies a common mathematical description [11–13]. Transport is usually described by an advection-dispersion equation, and the chemistry is formulated under thermodynamic equilibrium. A widely used approach to solve these reactive transport problems is the operator splitting approach [14]. Using this approach, the transport and chemical operators are solved separately at each time step and iteratively for some formulations. As a consequence, the chemistry operator has to be solved at least once per mesh cell per time step. This is one reason for the high computational cost of reactive transport modelling. Some authors have reported that 80 to 90 % of the computation time is dedicated to chemical computation. Many studies have been conducted to reduce the computation time required by reactive transport modelling [15]. Some works have explored parallelization [16], while others have focused on the methods used to solve the transport operator. Nevertheless, improving the resolution of the chemistry operator has been identified as a key

point. Some authors have attempted to improve the classic Newton-Raphson method [17], while others have tested other methods, such as Newton-Krylov [16, 18].

In this work, we focus on a specific element of the problem, improving the resolution of the linearized system provided by the Newton-Raphson method. Looking to numerical methods to solve linear systems is not currently a common practice. Indeed, these methods are actually well known [19–23], and all mathematical packages for scientific computation propose several routines for this task. The motivation of this work comes from the specificity of linear systems that have to be solved for equilibrium chemistry computations. Classic tests for the resolution of linear systems [24–30] are performed using systems provided by finite element or finite volume discretization, leading to matrices that are large (at least 10,000 unknowns) and sparse. Moreover, even when ill-conditioned systems have been studied [25, 30, 31], the conditioning of the matrix coming from the chemical system is specific, as underlined by Hoffmann et al. [32]. For example, Soleymani [33] worked with an ill-conditioned system constructed from 10×10 to 20×20 Hilbert matrices. The condition numbers then range from 3.5×10^{13} to 6.2×10^{28}. In this work, we present chemical tests leading to a 7×7 matrix with a condition number of approximately 10^{180}.

We expect to find a method to increase the efficiency of a speciation or reactive transport code. Several properties are required for such a method:

(i) This method should be fast, as the linear system will be solved very often. In the case of reactive transport modelling, the system will be solved at least once per mesh cell per time step.
(ii) The method should be very robust. It should be able to solve the linear system even if it is very poorly conditioned. Because the resolution of the linear system is only part of an iterative Newton step, an accurate solution is not absolutely needed. Thus, some advanced codes (e.g. Linear Algebra PACKage (LAPACK) routine) that check the accuracy of the solution and return an error flag instead of an inaccurate solution are, in this work, less robust than the more rustic routines.
(iii) The method should be able to detect failure and return an error flag to the main program so that a recovery procedure can be initiated. In the case of reactive transport modelling, this procedure could involve rejecting the current time step and recomputing with a smaller one.
(iv) In the initial analysis, the precision of the method is not the key point. Because the linear system resolution is only a part of the Newton-Raphson iterative procedure, *reasonable* error is acceptable for the linear system inversion. If this error is too large, it will slow the convergence speed for the Newton-Raphson method and decrease the efficiency of the reactive transport code. In this work, errors are estimated by comparing the calculated solution to a reference solution.

Because we utilize a markedly small matrix, we did not test parallelization. All the computations were performed on a PC running Windows with 64-bit Fortran 95. Real variables are defined as double-precision real. We prefer double-precision computations because all the chemical codes are, to the best of our knowledge, written as double-precision real and because quadruple-precision computation is much more time consuming. Nevertheless, we have tested one method using quadruple-precision real to determine whether this development could be useful. Reference solutions are also computed using quadruple precision.

We first present the formulation of the equations describing equilibrium reactions and how they are solved using the Newton-Raphson method. This point defines the Jacobian linear system, which is the object of this work. A second part is devoted to the presentation of the chemical tests and the numerical procedures used to perform the tests. Next, we propose a detailed analysis of the structure and properties of the Jacobian matrix. The selected linear solvers are then presented and tested, and the results are compared and discussed. Based on this analysis, we propose an algorithm to optimize the chemical computation in terms of robustness, accuracy, and efficiency. This algorithm is evaluated on the most selective test. By expanding the limits of the currently used methods, we believe that our new algorithm will contribute to enlarging the field of application of reactive transport modelling. As a conclusion, we underline the main advances of this work, the new perspectives and the remaining obstacles.

2 Material and methods

2.1 Geochemical modelling

One efficient formulation for the computation of thermodynamic equilibrium is based on the tableau concept, referred to as Morel's table [34, 35]. N_X components (X_j) are chosen from the N_C species (C_i) and are used to write the formation of each species as a combination of the components. The mass action law for the formation of the C_i species is written with the equilibrium constant (K_i) and the stoichiometric coefficients ($a_{i,k}$) for each component (X_k)

$$\{C_i\} = K_i \prod_{k=1}^{Nx} \{X_k\}^{a_{i,k}} \tag{1}$$

where $\{C_i\}$ and $\{X_k\}$ are the activities of species C_i and component X_k, respectively. In this work, we define X_j as a subset of C_i; then, N_X is N_C minus the number of reactions.

If N_{CP}-precipitated species (Cp_i) are taken into account, the mass action law for the precipitation of Cp_i is written with the precipitation constant (Kp_i) and the stoichiometric coefficients ($ap_{i,k}$). The saturation index (SI_i) of Cp_i is equal to its activity, which is unity for a pure solid phase

$$\mathrm{SI}_i = Kp_i \prod_{k=1}^{Nx} \{X_k\}^{ap_{i,k}} = 1 \tag{2}$$

The conservation of the total concentration $[T_j]$ of the jth component in the system is then written as

$$[T_j] = \sum_{i=1}^{N_C} a_{i,j} \cdot [C_i] + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i] \tag{3}$$

where $[C_i]$ is the concentration of species (C_i) and $[Cp_i]$ is the amount of precipitated species (Cp_i) per liquid volume unit.

A classic algorithm [17, 36–41] to describe mineral precipitation or dissolution makes an a priori hypothesis about the existence or non-existence of minerals. In this work, we assume that this hypothesis is proposed. The relationships between the activity and concentration are given by activity coefficients (γ_i) calculated using specific models (Davies, Debye-Hückel, etc.)

$$\{C_i\} = \gamma_i [C_i] \text{ and } \{X_j\} = \gamma_j [X_j] \tag{4}$$

By substituting the mass action law (1) into the mass conservation equation (3), the following relationship, which only depends on the components and the precipitated species concentrations, is obtained:

$$[T_j] = \sum_{i=1}^{N_C} a_{i,j} \cdot \left(\frac{K_i}{\gamma_i} \cdot \prod_{k=1}^{N_X} (\gamma_k [X_k])^{a_{i,k}} \right) + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i] \tag{5}$$

Combining Eqs. 2 and 5 leads to a set of $(N_X + N_{CP})$ nonlinear algebraic equations, which can be numerically solved through iterative methods. The concentrations of component $[X_k]$ and precipitated species $[Cp_i]$ at equilibrium are then determined when the $(N_X + N_{CP})$ objective functions (Y_j) are zero

$$\begin{aligned} Y_j &= -[T_j] + \sum_{i=1}^{N_C} a_{i,j} \cdot \left(\frac{K_i}{\gamma_i} \cdot \prod_{k=1}^{N_X} (\gamma_k [X_k])^{a_{i,k}} \right) + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i] \quad \text{for } j = 1 \text{ to } N_X \\ Y_{j=N_X+i} &= -1 + Kp_i \cdot \prod_{k=1}^{N_X} (\gamma_k [X_k])^{ap_{i,k}} \qquad \text{for } i = 1 \text{ to } N_{CP} \end{aligned} \tag{6}$$

Using this method, it is possible to include many chemical phenomena, including activity corrections, sorption on a surface using different means (such as ion exchange or surface complexation), and dissolution of gaseous compounds.

According to the criteria typically used for this method [17, 34, 40, 42], the convergence of the Newton-Raphson method is not checked with respect to the norm of the objective function $||Y||$, but the relative error defined as

$$\mathrm{NR}_{\text{relative error}} = \max \left[\left(\frac{|Y_j|}{|T_j| + \sum_{i=1}^{Nc} |a_{i,j} [C_i]|} \right)_{j=1,Nx}, \left(|Y_j| \right)_{j=Nx+1,Nx+NcP} \right] \leq \varepsilon_{N-R} \text{ with } \varepsilon_{N-R} = 10^{-12} \tag{7}$$

The value of the convergence criterion ($\varepsilon_{N-R} = 10^{-12}$) is set according to usual practice.

2.2 The Newton-Raphson method

The historical approach [12, 34, 37, 40, 42–47] involves the resolution of the system (6) with the Newton-Raphson method using $[X_k]$ and $[Cp_i]$ as primary unknowns. This formulation has some weaknesses that are explained later (see Section 3.1).

However, many authors [18, 32, 38, 39, 48] have proposed an alternative approach. Instead of using the component concentrations $[X_j]$ as the primary variables, they use the logarithm of the component activities ($\xi_j = \ln \{X_j\}$). According to this convention, the objective functions defined by Eq. 8 become conservation equations

$$Y_j = -[T_j] + \sum_{i=1}^{N_C} a_{i,j} \cdot \frac{K_i}{\gamma_i} \cdot \exp \left(\sum_{k=1}^{Nx} a_{i,k} \cdot \xi_k \right) + \sum_{i=1}^{N_{CP}} ap_{i,j} \cdot [Cp_i] \quad \text{for } j = 1 \text{ to } N_X \tag{8}$$

In the case of the objective function describing precipitation, it is more interesting to rewrite the mass action law (2) in log form and then define the objective function

$$Y_{Nx+i} = \ln(\mathrm{SI}_i) = \ln(Kp_i) + \sum_{k=1}^{Nx} ap_{i,k} \cdot \xi_k \text{for } i = 1 \text{to } N_{CP} \tag{9}$$

Equations 8 and 9 are solved at the nth iteration with the Jacobian matrix (Z^n) of the objective functions

$$\begin{aligned} Z^n_{j,k}\Big|_{\substack{j = 1, N_X + N_{CP} \\ k = 1, N_X}} &= \frac{\partial Y^n_j}{\partial [\xi_k]^n} \\ Z^n_{j,k}\Big|_{\substack{j = 1, N_X + N_{CP} \\ k = N_X + 1, N_X + N_{CP}}} &= \frac{\partial Y^n_j}{\partial [Cp_{k-N_X}]^n} \end{aligned} \tag{10}$$

Using an analytical computation, we obtain the $(N_X + N_{CP}) \times (N_X + N_{CP})$ values of Z^n by

$$\begin{aligned} Z^n_{j,k}\Big|_{\substack{j = 1, N_X \\ k = 1, N_X}} &= \sum_{i=1}^{N_C} a_{i,j} \cdot a_{i,k} \cdot [C_i]^n \\ Z^n_{j,k}\Big|_{\substack{j = 1, N_X \\ k = N_X + 1, N_X + N_{CP}}} &= ap_{k-N_X,j} \\ Z^n_{j,k}\Big|_{\substack{j = N_X + 1, N_X + N_{CP} \\ k = 1, N_X}} &= ap_{k,j-N_X} \\ Z^n_{j,k}\Big|_{\substack{j = N_X + 1, N_X + N_{CP} \\ k = N_X + 1, N_X + N_{CP}}} &= 0 \end{aligned} \tag{11}$$

Even if the activity coefficients depend on the component concentrations, they are assumed to be constant during the Newton-Raphson procedure. These activity coefficients are usually actualized by a fixed-point algorithm at each Newton-Raphson loop.

The progress step of the method $(\Delta\xi^n, \Delta Cp^n)$ is achieved by assuming that the objective function Y^{n+1} in Eq. 12 is equal to zero at the $(n+1)$th iteration. This produces the key equation of this article, the linear system (12), which must be solved to obtain the progress step

$$Z^n \cdot \left(\Delta\xi^n, \Delta Cp^n\right) = Y^{n+1} - Y^n = -Y^n \tag{12}$$

This system yields the values of the component activities and precipitate concentrations at the $(n+1)$th iteration

$$\begin{aligned} \xi^{n+1} &= \xi^n + \Delta\xi^n \\ [Cp]^{n+1} &= [Cp]^n + \Delta Cp^n \end{aligned} \tag{13}$$

To simplify the notations, ξ is used to denote the full vector of unknowns, including mineral Cp if present.

2.3 Chemical test cases

We choose chemical test cases with various numbers of components. Some of these chemical systems allow the formation of mineral species. Although it is not realistic from a chemical point of view, we test them without minerals and with the maximal possible number of minerals to obtain the largest matrix size. Appendix 1 presents the stoichiometric coefficients, equilibrium constants, and concentrations for these tests.

Appendices are avialable online.

(i) The *gallic acid* test case was presented by Brassard and Bodurtha [49]. It has been recognized as a challenging test for speciation computation [17] (see Appendix 1 (A-1)).

(ii) The *Valocchi* test is from Valocchi et al. [11]. It involves calcium and magnesium ion exchange (see Appendix 1 (A-2)).

(iii) The *pyrite* test case describes the dissolution of a pyrite rock in pure water. It has been used to test speciation algorithms [17]. Because it involves redox reactions, the stoichiometric coefficients cover a wide range, and the equilibrium constants vary over several orders of magnitude. This test is used under the assumption that no mineral phase is present (see Appendix 1 (A-3)).

(iv) The *MoMaS easy* test is the chemical system used for the reactive transport benchmark of MoMaS at the easy level [50]. It has been specifically developed to magnify numerical difficulties in a small system (see Appendix 1 (A-4)).

(v) The *Morel-Morgan* test is the first large chemical system reported in the computational literature. It was used by F. Morel and M. Morgan in 1972 to present the capacities of the computational method they had just developed (and which we still use today). This test includes 52 components (H^+, 20 metals, and 31 ligands), leading to 781 aqueous species (see Appendix 1 (A-5)).

(vi) The *MoMaS medium* test is the chemical system for the medium level of the MoMaS reactive transport benchmark [50] (see Appendix 1 (A-6)).

(vii) The *Fe-Cr* test is an additional redox test that describes the redox reactions between iron and chromium. These types of reactions occur when iron reactive barriers are used to treat chromium-contaminated sites [51, 52]. In this case, we consider only the aqueous phase without minerals (see Appendix 1 (A-7)).

(viii) The *pyrite mineral* test describes the dissolution of a pyrite rock in pure water. We assume that three possible mineral phases are present (see Appendix 1 (A-8)).

(ix) The *MoMaS hard* test is the equilibrium part of the chemical system described in the hard level of the MoMaS reactive transport benchmark. It allows for the formation of two mineral species (see Appendix 1 (A-9)).

(x) The *Fe-Cr mineral* test describes the redox reaction between iron and chromium. We assume the formation of three different mineral phases (see Appendix 1 (A-10)).

2.4 Test procedure

Equation 11 shows that we can obtain multiple linear systems from one chemical problem by changing the activity values of the components. For each chemical system, we select three components and vary their values over a wide range. The concentrations of all minerals are arbitrarily set to 10^{-3} mol L^{-1}. The activity of component H^+ is varied from 10^{-12} to 10^{-2} mol L^{-1} (pH = 12 to pH = 2), while that of component e^- is varied from 10^{-19} to 10^{12}, corresponding to Eh = −0.7 to 1.1 V computed using Eq. 14 at 25 °C

$$\mathrm{Eh} = \ln\left\{e^-\right\}\frac{RT}{F} \tag{14}$$

where T is the temperature (Kelvin), R is the gas constant (8.314 J K mol^{-1}), and F is the Faraday constant (96,487 C mol^{-1}). This range of electrical potential corresponds to the stability of water at pH values between 2 and 12. For the O_2 component, it is not possible to cover the same potential range as e^- because of the computation of the reference solution. The activity is varied from 10^{-70} to 10^{4}, as computed using Eq. 15 at 25 °C with $E^0 = 1.23$ V and pH varying from 2 to 12. The potential is then varied from −0.5 to 1.1 V

$$\mathrm{Eh} = E^0 + \frac{1}{4}\frac{RT}{F} \times \ln\frac{\{\mathrm{O_2}\}\left\{\mathrm{H^+}\right\}^4}{\{\mathrm{H_2O}\}} \tag{15}$$

The activities of the other components vary from 10^{-12} to 10^{-1} mol L^{-1}. For each of the three selected components, we compute 30 values equally distributed on a log scale over the chosen range, leading to 29,791 different linear systems for each chemical test case. For each of these 29,791 tests, we make only one linear solver (or one Newton step) (except in the last section, Section 4, where the iterative Newton method is performed to solve the non-linear system given by Eqs. 8 and 9).

The matrix norm used in this work is the $||-||_1$ norm, defined as [23]

$$\|Z\|_1 = \max_{1\le j\le n}\left(\sum_{i=1}^{n}\left|Z_{i,j}\right|\right) \tag{16}$$

The condition number of Z is defined [23] as the product of the norm of the matrix per the norm of the inverse matrix (17)

$$\mathrm{cond}\,(Z) = \|Z\|_1 \times \left\|Z^{-1}\right\|_1 \tag{17}$$

To test the numerical methods, we first evaluate the computation time (CPU time) required to solve the linear system. Because we work with a very small matrix, the computations are very fast and we run the same calculation several times to obtain a total computing time of approximately 1 s. The *CPU time* is given in this work in units of seconds per computation (by dividing the total computing time by the number of runs). According to this method, the global computing time for one test case is approximately 6 days.

Many numerical methods, including a *failure indicator*, which indicates the success or failure of the resolution, have been developed. If needed, we include a failure indicator. As *failure*, we include the *crash* of the method, underflow or overflow, non-convergence within the maximum number of iterations (for iterative methods), or excessive inaccuracy for some advanced methods (LAPACK routines) that estimate the accuracy of the proposed solution.

Solving a linear system (13) using a numerical method produces an approximate solution ($d\xi_{\mathrm{method}}$), and the reference method gives ($d\xi_{\mathrm{ref}}$) with accuracy on the same order as the roundoff error. To evaluate the accuracy of the approximate solution, two quantities can be calculated:

1. The *relative error on the norm*, $\mathrm{Err_{Norm}}$, is obtained by computing the norm of the approximate and reference solution (18)

$$\mathrm{Err_{Norm}} = \frac{\left|\|d\xi_{\mathrm{method}}\| - \|d\xi_{\mathrm{ref}}\|\right|}{\|d\xi_{\mathrm{ref}}\|} \tag{18}$$

1. The error on the direction is given by $angle_{method}$, the angle (degrees) between the reference and the approximate solution calculated using the scalar product of these two vectors

$$\mathrm{angle_{method}} = \frac{360}{2\pi}Arc\cos\left(\frac{d\xi_{\mathrm{method}}\cdot d\xi_{\mathrm{ref}}}{\|d\xi_{\mathrm{method}}\|\cdot\|d\xi_{\mathrm{ref}}\|}\right) \tag{19}$$

All of these quantities, namely the failure indicator, relative error on the norm, $\mathrm{angle_{method}}$, and CPU time, are calculated for the 29,791 linear systems built from each chemical test case for all the tested methods. This enormous amount of data is aggregated in two ways:

(i) For each chemical system and each method, we compute the mean of each quantity.

(ii) For each chemical system and each method, the interval of the condition number is discretized into 100 regular subintervals. For each subinterval, we compute the mean of each quantity.

2.5 Reference solution

Because of the very high condition numbers, it is not possible to directly obtain an exact solution. We equilibrate the rows and columns of the Jacobian matrices to reduce their condition number using the iterative algorithm proposed by Knight et al. [53] because it preserves the symmetry of the Jacobian matrix.

Let $\widetilde{\mathbf{Z}}^k$ be the equilibrated Jacobian matrix at iteration k, with $\widetilde{\mathbf{Z}}^0 = Z$.

These authors defined r_i^k as the vector formed by the ith row of $\widetilde{\mathbf{Z}}^k$ and c_i^k as the vector formed by the ith column. The preconditioning matrices $\mathbf{R}^k$ and $\mathbf{C}^k$ are then defined by

$$\mathbf{R}^k = \mathrm{diag}\left(\frac{1}{\sqrt{\|r_i^k\|_\infty}}\right)_{i=1,Nx+NcP} \quad \text{and} \quad \mathbf{C}^k = \mathrm{diag}\left(\frac{1}{\sqrt{\|c_i^k\|_\infty}}\right)_{i=1,Nx+NcP} \tag{20}$$

The equilibrated matrix is defined at iteration $k+1$ by

$$\widetilde{\mathbf{Z}}^{k+1} = \mathbf{R}^k \cdot \widetilde{\mathbf{Z}}^k \cdot \mathbf{C}^k \tag{21}$$

This procedure is repeated until all $\|r_i^k\|_\infty$ and $\|c_i^k\|_\infty$ are equal to 1 or after 50 iterations. Let $\mathbf{R}$ and $\mathbf{C}$ be the resulting preconditioning matrices and $\widetilde{\mathbf{Z}}$ the equilibrated matrix. Instead of solving the linear system (12), we solve

$$\widetilde{\mathbf{Z}} \cdot \widetilde{\mathbf{x}} = -\widetilde{\mathbf{Y}} \tag{22}$$

where $\widetilde{\mathbf{x}} = \mathbf{C}^{-1} \cdot (\Delta\xi, \Delta Cp)$ and $\widetilde{\mathbf{Y}} = \mathbf{R} \cdot Y$. These procedures are coded using quadruple-precision reals. The linear system (22) is solved by LU decomposition coded with quadruple-precision real.

Even if the condition numbers of the Jacobian matrices (Z) are very high ($10^{213.9}$ for the Fe-Cr mineral test case), the condition numbers of the equilibrated matrices ($\widetilde{\mathbf{Z}}$) are much lower: the maximum condition number obtained after equilibration is $10^{13.4}$. According to Golub and van Loan [54], if the unit roundoff is approximately 10^{-d} and the condition number is approximately 10^q, then the Gaussian elimination gives a solution with approximately $d - q$ correct digits. Because we use quadruple precision, we obtain $d = 32$, leading to $32 - 14 = 18$ correct digits. One can then assume that the reference solution is exact if we compare it to the solutions produced by the tested methods (computed using double-precision real).

2.6 Selected numerical methods for solving linear systems

Studies on linear algebra [19, 23] present methods for solving linear systems as direct or iterative methods. Historically, speciation codes solved linear systems using direct methods, such as Gaussian elimination [34] or LU decomposition [17, 40, 42]. In its actual form, the speciation code SPECY [48] uses unsymmetric multifrontal (UMF) [55] as the linear solver. To the best of our knowledge, no speciation code uses iterative methods to solve linear systems. This point is in accordance with the existing literature, which reports the use of iterative methods for solving large, sparse linear systems [20–22, 24, 26, 28, 29, 56, 57]. Nevertheless, actual developments in speciation codes involve the use of large chemical databases [39, 58, 59], leading to an increase in the size of the chemical systems. The use of iterative methods is also studied in this work.

We select some direct and iterative solvers according to the properties of the linear systems and the speciation computation methods currently in use (Table 1).

For the direct method, we select LU decomposition [60] because it was originally used for speciation computations by Westall [40] and Westall et al. [42]. The UMF method [55] has been implemented in the speciation code SPECY [48] in place of the LU approach [17]. After showing that the Jacobian matrix is symmetric, we test the DSYTRS subroutine from LAPACK [61], which is based on a UDU decomposition. Because the Jacobian matrix is often positive definite, as shown in Table 3, we test the DPOTRS subroutine [61] based on the Cholesky method. Some authors [32] have used iterative QMRCGStab to solve reactive transport under a global approach. Here, we test QR decomposition using the DGELS routine [61].

For the iterative methods, we test the Jacobi [23, 62], Gauss-Seidel [23, 62], and successive over-relaxation (SOR) [23, 62] methods. Barrett et al. [21] proposed an algorithm to select an iterative solver depending on the matrix properties. GMRES was presented as the least selective algorithm. We use a GMRES method developed by HSL [63]. If the matrix is symmetric, Barrett et al. [21] recommend the use of conjugate gradient squared (CGS) or biconjugate gradient stabilized (BiCGStab) methods. CGS and BiCGStab subroutines have been developed by HSL. We test two additional methods devoted to symmetric matrices: SYMMBK [63] and an incomplete Cholesky (Inc. CHOLESKY) factorization [63].

We use the same parameters for all iterative methods: a maximum of 500 iterations and a stopping criterion of 10^{-8}. To determine the influence of the stopping criterion, we test the GMRES method using 50,000 maximum iterations and 10^{-12} as the stopping criterion, denoted by GMRES 10^{-12} in this study. A critical point of the GMRES algorithm is the size of the Hessenberg matrix. In this work, we set it to the max of 8 (Nx + NcP).

The results obtained using the Jacobi and SOR methods are not detailed here. As previously reported [19], the Jacobi method is inefficient, leading to a very high failure ratio (close to 100 %) even for the easiest test cases. For the SOR

Table 1 List of the selected solvers

Name	Source	Method	Matrix properties
Direct			
LU	[60]	LU decomposition	–
DGETRS	[61]	LU decomposition	–
UMF	[55]	Direct multifrontal	–
DSYTRS	[61]	UDU-factored symmetric matrix	Symmetric
DPOTRS	[61]	Cholesky $A = U^T \times U$	Definite positive
DGELS	[61]	QR decomposition	
LU QUAD	[60]	LU decomposition quadruple precision	–
Iterative			
SYMMBK	[63]	Iterative SYMMBK HLS_MI02	Symmetric
Inc. CHOLESKY	[63]	Incomplete Cholesky HSL_MI28	Symmetric
CGS	[63]	Conjugate gradient squared HLS_MI23	–
BiCGStab	[63]	Biconjugate gradient squared stabilized HLS_MI26	–
GMRES	[63]	Flexible GMRES HLS_MI15	–
Gauss-Seidel	[60]	Gauss-Seidel method	–
Preconditioned			
LU Equil	[53–60]	LU and matrix equilibration	–
DGESVX	[61]	LU and optional preconditioning	–
GMRES Equil	[53–63]	GMRES and matrix equilibration	–
GMRES 1.d-12	[63]	GMRES convergence criteria 1.d-12	–

method [23, 26, 56, 62], the over-relaxation parameter is the key factor. Unfortunately, we did not find any efficient relationships to define it. For the same chemical system, the best value varies from 0.097 to 1.91 without apparent order.

We do not extensively test the possibility of using a preconditioner. As stated by Barrett et al. [21]: "Since using a preconditioner in an iterative method incurs some extra cost, both initially for the setup, and per iteration for applying it, there is a trade-off between the cost of constructing and applying the preconditioner, and the gain in convergence speed". In our case, the matrices are very small, leading us to suppose that this trade-off would not be advantageous. Nevertheless, an easy way to test preconditioners is proposed by the LAPACK routine DGESVX, which performs LU decomposition and matrix equilibration depending on the estimated condition number. We implement matrix equilibration according to Knight et al. [53] to obtain a reference solution. We test this preconditioning technique associated with LU decomposition and the GMRES method, denoted by LU Equil and GMRES Equil in this study. The maximum iterations allowed for the equilibration procedure is fixed to 5, according to the recommendations of Knight et al.

Table 2 Structure of the Jacobian matrix

Z	$\frac{\partial}{\partial \xi_k}$	$\frac{\partial}{\partial [Cp_{k-N_x}]}$
$\frac{\partial Y_j}{\partial}$	$\sum_{i=1}^{N_C} a_{i,j} \cdot a_{i,k} \cdot [C_i^n]$ (above diagonal); $\sum_{i=1}^{N_C} a_{i,j} \cdot a_{i,k} \cdot [C_i^n]$ (below diagonal)	$ap_{k-N_x,j}$
$\frac{\partial Y_j}{\partial}$	$ap_{k,j-N_x}$	0

Table 3 Properties of the 10 chemical test cases ranked by increasing the maximal condition number

	Nx	Nc	NcP	Z size	cond(Z) min	cond(Z) max	cond(Z) max after 50 equilibration	%Z diagonal dominant	%Z positive definite
Gallic acid	3	17	0	3	$10^{0.61}$	$10^{12.6}$	$10^{0.95}$	18.4	100
Valocchi	5	7	0	5	$10^{0.49}$	$10^{15.3}$	$10^{0.65}$	67.7	100
Pyrite	4	40	0	4	$10^{4.06}$	$10^{24.9}$	$10^{0.95}$	0.00	100
MoMaS easy	5	12	0	5	$10^{3.44}$	$10^{37.7}$	$10^{1.05}$	0.00	71.1
Morel-Morgan	52	781	0	52	$10^{43.4}$	$10^{60.7}$	$10^{1.13}$	0.00	35.9
MoMaS medium	5	14	0	5	$10^{5.88}$	$10^{103.9}$	$10^{0.95}$	0.00	78.8
Fe-Cr	7	39	0	7	$10^{9.46}$	$10^{113.6}$	$10^{1.05}$	0.00	68.9
Pyrite mineral	4	43	3	7	$10^{1.71}$	$10^{33.1}$	$10^{3.19}$	0.00	0.00
MoMaS hard	6	15	2	8	$10^{5.45}$	$10^{123.9}$	$10^{3.02}$	0.00	0.00
Fe-Cr mineral	7	43	3	10	$10^{8.67}$	$10^{213.9}$	$10^{13.4}$	0.00	0.00

Finally, we test an LU decomposition method compiled as quadruple precision, denoted by LU QUAD. The source of this method is the LU double-precision real of numerical recipes [60], and we adapt it to quadruple precision. Because the usual computations are performed using double precision, the quadruple precision ($d\xi_{\mathrm{QUAD}}$) should be translated in double-precision real. To avoid overflow, we rescale $d\xi_{\mathrm{QUAD}}$ to ensure its validity. If huge $(1.d0)$ is the highest double-precision real represented by the machine, we rescale $d\xi_{\mathrm{QUAD}}$ to obtain the double-precision solution $d\xi_{\mathrm{LU\ QUAD}}$:

$$d\xi_{\mathrm{LU\ QUAD}} = \frac{\mathrm{huge}\ (1.d0)}{\max\left|d\xi_{\mathrm{QUAD}}\ (i)\right|} \cdot d\xi_{\mathrm{QUAD}} \tag{23}$$

In this way, we conserve the direction of the Newton step, even if its norm is changed.

3 Results and discussion

3.1 Properties of the Jacobian matrices

As defined by Eq. 11, the Jacobian matrix has several properties:

(i) The matrix is block-structured, as presented in Table 2. A four-block structure is present if precipitation occurs.

(ii) The matrix is symmetric, as shown in Table 2.

(iii) In the case of no precipitation, all the diagonal terms of the matrix are strictly positive because they are the sum of $a_{i,j}^2\,[C_i]$. It is then possible for the matrix to be diagonal dominant. We examine this possibility for the selected test case. Table 3 shows the ratio of diagonal dominant Jacobian matrices for all the chemical tests performed according to the previously defined test procedure. Some matrices in the gallic acid and Valocchi cases are diagonal dominant, but none of the matrices from the other cases are diagonal dominant. By plotting the ratio of diagonal dominant matrices depending on the condition number (see Appendix 2 (B-1)), it appears that only matrices with very low condition numbers can be diagonal dominant.

(iv) Because the Jacobian matrix is real, symmetric, and sometimes diagonal dominant, the question of whether it is positive definite may be posed. In the case of no precipitation, Eq. 11 can be written in matrix form, leading to Eq. 24

$$Z = A^T \cdot \mathrm{diag}\ (C) \cdot A \tag{24}$$

Because the concentrations are positive, the Jacobian matrix is analytically positive definite. Nevertheless, this may not be true numerically. We are not able to propose a general framework, but we can compute the eigenvalues of the Jacobian matrix and test whether they are positive for all test cases. Table 3 shows that for the gallic acid, Valocchi, pyrite, and Morel-Morgan test cases, all the Jacobian matrices are positive definite. For the MoMaS easy, MoMaS medium, and Fe-Cr test cases, a large proportion (66.4 to 74.1 %) of the Jacobian matrices are positive definite. For cases including minerals (pyrite mineral, MoMaS hard, and Fe-Cr), essentially none of the matrices are positive definite (only 0.1 % for the MoMaS hard test). Plotting the ratio of positive definite matrices as a function of the condition number (see Appendix 2 (B-2)) shows that the chemical conditions are more important than the condition number when determining whether the Jacobian matrix is diagonal dominant.

(v) According to the test procedure presented previously, we plot, on the same graph, the logarithm of the norm

of $||Y||$ and the logarithm of the condition number of the matrix Z (Fig. 1). There is a strong linear relationship between these parameters. Moreover, the linear relationship does not depend on the chemical test, only on the existence of minerals. According to our results, the conditioning of the Z matrix can be evaluated using the following empirical formulas:

$$\begin{aligned} \text{cond}\,(Z)_{\text{no mineral}} &= 10^{5.30\pm0.03} \times \|Y\|^{0.9374\pm0.0008} \\ \text{cond}\,(Z)_{\text{mineral}} &= 10^{-3.23\pm0.08} \times \|Y\|^{1.706\pm0.002} \end{aligned} \tag{25}$$

The value and uncertainties are obtained through the least squares method over all cond(Z) and $||Y||$. In this way, we propose an estimation of cond(Z) with no computation time cost because the objective function is evaluated during the Newton-Raphson procedure. As shown in Fig. 1, cond(Z) and $||Y||$ are strongly correlated for large condition numbers, and the results are noisier if cond(Z) and $||Y||$ are small. The evolution of this relation for low $||Y||$ can be seen in Appendix 7 (G-11). Therefore, Eq. 25 should not be used for $||Y||$ less than 10^{10}.

Several of these properties are obtained using the logarithm of the component activities as the primary unknown in Eq. 8. The historical approach [34] uses the component concentrations as the primary variable and leads to a less interesting Jacobian matrix. Even if the structure presented in Table 2 exists, the matrix is not symmetric. Moreover, the matrix is worse conditioned (condition number from $10^{11.2}$ to $10^{49.4}$ rather than $10^{4.06}$ to $10^{24.9}$ for the pyrite case). Finally, no specific relation exists between cond(Z) and $||Y||$ for the historical formulation.

As an example, we show one linear system from the Fe-Cr mineral test, corresponding to a condition number of 10^{187}. One can observe the structure of the matrix and the specificity of the linear system (26).

$$\begin{bmatrix} 1.15\cdot10^{94} & 9.09\cdot10^{93} & -5.04\cdot10^{-13} & -11.7 & 3.03\cdot10^{93} & 0 & 1.10\cdot10^{87} & 0 & 5 & -1 \\ & 5.45\cdot10^{93} & 0 & 1.14\cdot10^{10} & 1.82\cdot10^{93} & 0 & 4.11\cdot10^{86} & 2 & 3 & 0 \\ & & 1.00\cdot10^{-6} & 2.37\cdot10^{-15} & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 2.91 & 8.74 & 0 & 1.28\cdot10^{-6} & 1 & 0 & 0.25 \\ & & & & 6.06\cdot10^{92} & 2.23\cdot10^{-2} & 1.37\cdot10^{86} & 0 & 1 & 0.75 \\ & & & & & 2.85\cdot10^{-2} & 1.71\cdot10^{-4} & 0 & 0 & 0 \\ & & & & & & 1.37\cdot10^{86} & 0 & 0 & 0 \\ & & & & & & & 0 & 0 & 0 \\ & & & & & & & & 0 & 0 \\ & & & & & & & & & 0 \end{bmatrix} \cdot (d\xi) = \begin{pmatrix} -3.03\cdot10^{93} \\ -1.82\cdot10^{94} \\ -1.05\cdot10^{-13} \\ 8.99\cdot10^{-3} \\ -6.06\cdot10^{94} \\ -2.25\cdot10^{-2} \\ -1.37\cdot10^{86} \\ -27.6 \\ 180 \\ 3.84 \end{pmatrix} \tag{26}$$

3.2 Robustness of the methods

Figure 2 presents the failure ratio for each method and each test case. The presence of minerals prevents the DPOTRS, Inc. CHOLESKY, and Gauss-Seidel methods from solving the system. If there are minerals present in the chemical system, a zero-value block appears in the Jacobian matrix, as shown in Table 2 and Eq. 26. This block makes the Inc. CHOLESKY factorization unappropriated. Because the Gauss-Seidel method requires division by each diagonal term, this zero-value block makes the method unadapted. The failure of the DPOTRS routine is explained by the properties of the Jacobian matrix. As shown in Table 3, there is no positive definite matrix in the presence of minerals. In the case of the DPOTRS, Inc. CHOLESKY, and Gauss-Seidel methods, the term failure is ambiguous. These methods are *expected to fail* and should not be used on systems with minerals. If there are no minerals, some matrices are not positive definite in the MoMaS easy, MoMaS medium, Morel-Morgan, and Fe-Cr tests. This explains the failure of the DPOTRS routine.

Some other methods (DGETRS, DSYTRS, DGELS, and DGESVX) present a substantial failure ratio, mainly for high condition number tests (MoMaS easy and Fe-Cr mineral). UMF, SYMMBK, and CGS are robust for the Fe-Cr mineral test but present significant failure ratios for lower-conditioned tests, such as MoMaS easy or pyrite mineral. Some methods adapted to symmetric matrices (DSYTRS and SYMMBK) are included in this class of weak methods.

The BiCGStab method has a very low failure ratio and fails only in the two difficult tests (MoMaS easy and Fe-Cr mineral). GMRES is the only successful iterative method.

Figure 2 shows that some methods are successful for all the test cases. The most successful direct method is LU, while the most successful iterative methods are GMRES and GMRES 10^{-12}. The quadruple-precision method LU QUAD is also successful, which is expected because the double-precision LU method is also successful. The use of an equilibration method as a preconditioner makes LU Equil and GMRES Equil successful.

As stated previously, we focus on the capacity of a method to produce a solution independent of its accuracy.

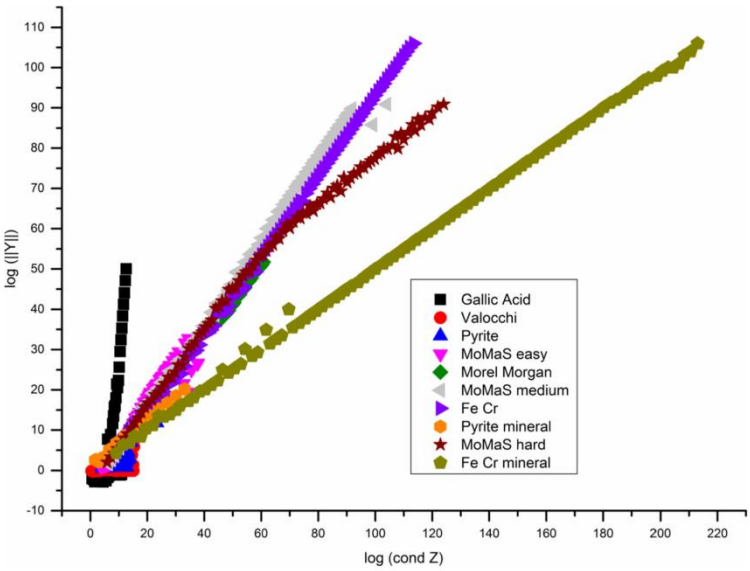


Fig. 1 Relationship between the condition number of Z and the norm of the objective function plotted on a log-log scale

For some advanced methods (e.g. LAPACK methods), a posteriori estimation of the residual and estimation of the condition numbers are performed. If the solution is not sufficiently accurate, no solution is given, leading to a higher failure ratio than for the more rustic methods (LU or Gauss-Seidel). Because the key point of this work—the resolution of a linear system—is included in the iterative Newton procedure, it is preferable to obtain an inaccurate solution (so

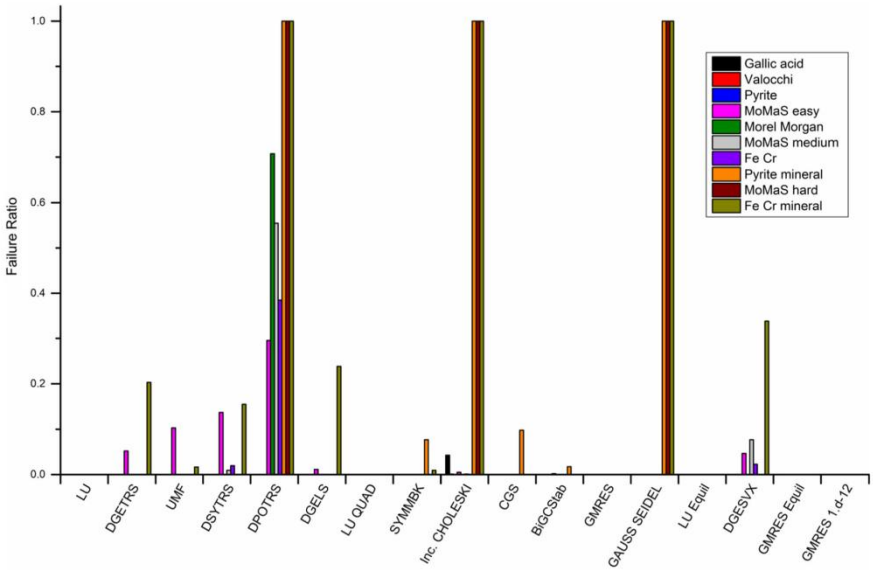


Fig. 2 Mean of the failure ratio for each method and each test case

the iterative procedure can be continued) than no solution (the iterative procedure will be aborted).

Appendix 3 presents the evolution of the failure ratio for each test case and each method depending on the condition number.

For the direct methods (Appendix 3 (C-1 to C-5)), for small condition numbers corresponding to the test cases gallic acid, Valocchi, and pyrite, no failure occurs. As the condition number increases, the failure ratio also increases for some methods. MoMaS easy (Appendix 3 (C-4)), MoMaS medium (Appendix 3 (C-6)), and Fe-Cr (Appendix 3 (C-7)) show that for condition numbers greater than 10^{20}, the failure ratio increases greatly for some of the methods. These methods are DOPTRS and DSYTRS for MoMaS medium and Fe-Cr. DGETRS, UMF, DSYTRS, DOPTRS, and DGELS present some failure for condition numbers greater than 10^{15} for the MoMaS easy case. In the presence of minerals (Appendix 3 (C-8 and C-9)), for low condition numbers (the pyrite mineral case), the methods are either successful (UMF, LU, DSYTRS, DGETRS) or completely unsuccessful (DPOTRS). For very high condition numbers (Fe-Cr mineral case), the success of the method does not depend on the condition number. We suppose that the condition numbers (see Table 3) are too high to exhibit any ordering.

For other iterative methods, the success does not depend on the condition number but on the nature of the matrix and the presence (Appendix 3 (C-18 to C-20)) or absence (Appendix 3 (C-11 to C-17)) of minerals.

3.3 Accuracy of the methods

The accuracy of the methods is evaluated in two ways: (i) the relative error on the norm (18) and (ii) the angle between the reference and the calculated solution (19).

(i) By plotting the mean of the logs of the relative error on the norm of each test case (Fig. 3), some general tendencies are identified. The relative residual tends to increase with the condition number of the system. For direct methods and small condition numbers, the relative residual is small (10^{-10} to 10^{-3}) for the gallic acid, Valocchi, and pyrite test cases. For the iterative methods, the relative residual corresponding to an accurate resolution for tests with small condition numbers is approximately 10^{-4}. This value corresponds to the value of the convergence criteria of the iterative methods. Iterative methods are more sensitive to the condition number than direct methods. Only the Valocchi test case is accurately solved by almost all the iterative methods, whereas the first three tests are accurately solved by all the direct methods. Even in the case of successful resolution (CGS and BiCGStab methods), the relative errors on the norm are high for intermediate cases (pyrite, MoMaS easy, and Morel-Morgan). Nevertheless, the results are better for the iterative methods than for the direct methods for the difficult tests (MoMaS easy, MoMaS medium, MoMaS hard, Fe-Cr mineral). The GMRES and Gauss-Seidel methods have mostly constant mean relative error on the norm, with the same accuracy for all test cases. GMRES and Gauss-Seidel are less efficient than the other methods for the easy tests, but more ill-conditioned tests are better solved by these two methods.

The condition numbers are so high that even LU QUAD cannot provide accurate resolution. For the MoMaS medium and Fe-Cr mineral tests, many of the solutions calculated by the LU QUAD method are rescaled using Eq. 23, leading to excessively high relative error on the norm.

Comparison of the relative error on the norm given by the non-preconditioned (LU, DGETRS, and GMRES) and preconditioned (LU Equil, DGESVX, and GMRES Equil) methods shows that the preconditioned methods lead to lower relative error than the non-preconditioned methods for the direct methods, but the result is more case-dependent for GMRES. The use of preconditioning usually leads to lower relative error on the norm, except for the Morel-Morgan, Fe-Cr, and MoMaS hard cases.

Increasing the maximum number of iterations and reducing the convergence criteria of GMRES leads to less relative error on the norm, but this reduction is not significant.

Nevertheless, the global means of the logs of relative errors on the norm hide the influence of the increasing condition number. Appendix 4 presents the evolution of the relative error on the norm for each test case and each method depending on the condition number. The theoretical behaviour is verified for the direct methods and for all the test cases (except for the Valocchi one, Appendix 4 (D-2)). The relative error on the norm increases regularly with the condition number. It is close to 10^{-16} when the condition number is close to 1 and increases to 1 when the condition number is close to 10^{16}, in accordance with the computation theory presented by Golub and van Loan [54]. For condition numbers greater than 10^{16}, the evolution of the relative error on the norm with the condition number is much noisier. The use of the quadruple-precision LU QUAD method leads to an accurate resolution of a large portion of the tested systems. As expected by computation theory, all the systems with condition numbers less than 10^{32} are solved with a relative error on the norm of approximately 10^{-15}. In some cases (MoMaS medium (Appendix 4 (D-6)), Fe-Cr (Appendix 4 (D-7)), MoMaS hard (Appendix 4 (D-9))), LU QUAD produces an increasing relative error

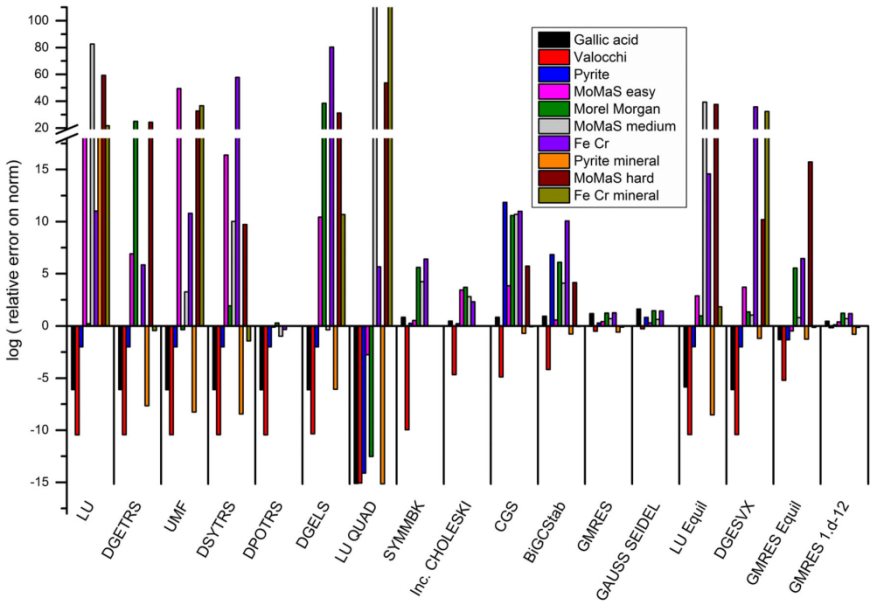


Fig. 3 Mean of the logs of the relative error on norm for each method and each test case

with increasing condition number (if higher than 10^{32}) but not systematically. LU QUAD produces a very low relative error on the norm even if the condition number is very high (Appendix 4 (D-9)). This behaviour can be explained by the fact that the LU QUAD method and/or the reference method is unable to exactly solve such ill-conditioned systems. LU QUAD produces a very high relative error on the norm, one point with 10^{290} error for the MoMaS medium (Appendix 4 (D-6)), and all the values at condition numbers greater than 10^{90} for the Fe-Cr mineral (Appendix 4 (D-10)) test case. These points correspond to the rescaling of the computed quadruple-precision solution to maintain it on the double-precision scale (using Eq. (23)).

Iterative methods present similar behaviour to direct methods, giving very low relative error on the norm (between 10^{-15} and 10^{-8}) when the condition number is less than a critical value. This critical value depends on the method and the test case. It can be set to 10^8 for SYMMBK CGS, BiCGStab, and GMRES for the gallic acid (Appendix 4 (D-11)) and MoMaS easy (Appendix 4 (D-14)) cases. It can be set to 10^{12} or 10^{15} for Inc. CHOLESKY for the gallic acid and MoMaS easy cases and for SYMMBK, Inc. CHOLESKY, CGS, BiCGStab, and GMRES for the pyrite (Appendix 4 (D-13)), the MoMaS medium (Appendix 4 (D-16)), and MoMaS hard (Appendix 4 (D-19)) tests. Using low convergence criteria (GMRES 1.d-12) leads to lower relative error on the norm for low condition numbers (Appendix 4 (D-21 to D-23, D-26 to D-29)), but no significant improvements are obtained if the condition number increases, as shown in Appendix 4 (D-24 to D-30).

Using preconditioning methods reduces the relative error on the norm for intermediate condition numbers. No gain is obtained for low condition numbers (Appendix 4 (D-21 and D-22)), but the errors given by LU Equil, DGESVX, and GMRES Equil are less than the LU and GMRES errors for higher condition numbers (Appendix 4 (D-24 to D-26)). For very high condition number tests (Appendix 4 (D-27, D-29, and D-30)), the errors given by the preconditioned methods are equivalent to the errors given by the non-preconditioned methods.

(ii) By plotting the angle between the reference solution and the calculated solution, we can compare the methods according to the computed direction (Fig. 4). Because the resolution of the linear system (13) represents one step in the iterative Newton procedure, this information is much more important than the norm of the step. A wrong norm can be corrected using line search methods [64], whereas modifying a wrong direction leads to additional iterations. Small condition number tests (gallic acid, Valocchi, pyrite, and pyrite mineral) are solved using direct methods with the right direction. If the condition number

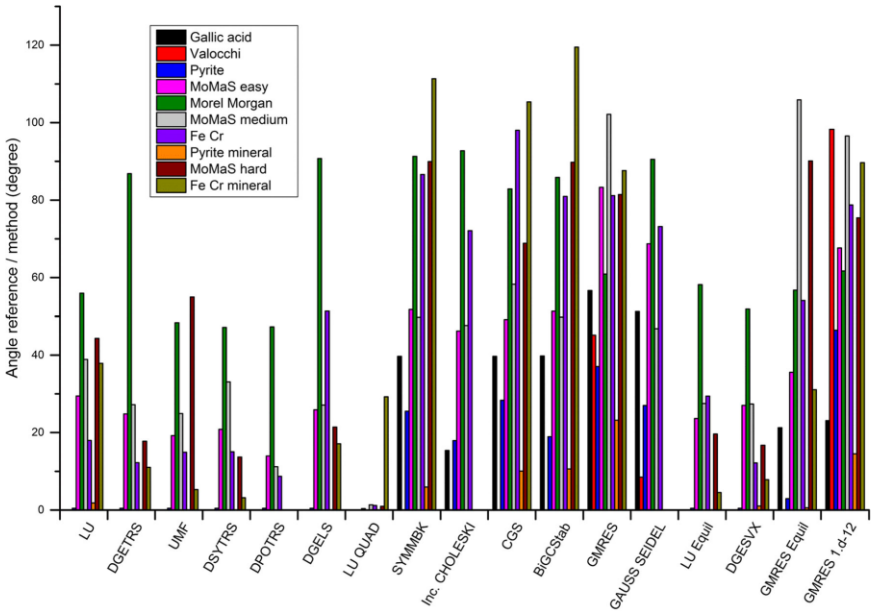


Fig. 4 Mean of the angles between reference and computed solution for each method and each test case

increases, the directions given by the direct methods become inaccurate, but the condition number is not the only governing parameter. Morel-Morgan leads to worse direction than MoMaS medium and Fe-Cr, and MoMaS hard leads to a higher angle than the Fe-Cr mineral test. Iterative methods result in a worse direction than direct methods, and only the Valocchi test case is solved with an accurate direction by all the iterative methods. Imposing lower convergence criteria (10^{-12}) on GMRES leads to a worse direction than using the usual criteria (10^{-8}). Using preconditioning methods leads to a better direction when associated with a direct method (LU Equil and DGESVX), but the conclusion is less clear for the iterative GMRES Equil method. Depending on the test case, the direction can be worse (Valocchi, MoMaS easy, MoMaS medium) or better (gallic acid, pyrite, MoMaS hard, Fe-Cr mineral)

The influence of the condition number on the angle (see Appendix 5) indicates that the direction is correct for direct methods when the condition number is less than 10^{15}. For iterative methods, the limit to obtain an accurate direction is a condition number less than 10^{8}, excepted for the Gauss-Seidel method, which produces wrong directions for low condition numbers. If the condition number increases, the behaviour of the direction becomes noisy. Since the relative error on the norm increases regularly until the condition number reaches the limit of 10^{8} or 10^{15}, the angle is accurately defined until this condition number limit is reached. Using preconditioned methods leads to a better direction for the LU Equil and the GMRES Equil methods when the condition number is higher than 10^{15} for some cases (Appendix 5 (E-21, E-23 to E-25, and E-30)) but to a worse direction for other cases (Appendix 5 (E-26 and E-29)).

We present two successful direct methods, LU and LU QUAD; one iterative method, GMRES (both tested versions, GMRES and GMRES 10^{-12}); and two preconditioned methods, LU Equil and the GMRES Equil. By comparing the relative error on the norm (Appendix 5 (D-21 to D-30)), the successful methods can be ranked from the lowest to highest error: LU QUAD, GMRES 10^{-12}, GMRES Equil, LU Equil, and LU. Ranking these methods according to the angle between the reference and computed solution is more complicated. For all the tests cases (Appendix 5 (E-21 to E-25, E-27, E-28, and E-30)), LU QUAD gives the best direction, followed by LU Equil, LU, GMRES Equil, and GMRES 10^{-12}. The MoMaS medium (Appendix 5 (E-26)) and MoMaS hard (Appendix 5 (E-29)) test cases lead to the same conclusion, except GMRES Equil which gives the worst direction.

3.4 Efficiency of the methods

The speed of the methods is studied by recording the computation time for each test case and plotting the mean CPU time for each test case and each method (see Fig. 5). As expected, the computation times are very short (less than 1 ms) because the systems to solve are small.

Figure 5 shows the influence of the system size. For all methods, the computation time increases with the number of unknowns. The results show that the iterative methods are less sensitive to the system size than the direct methods. For the iterative methods, the number of iterations is important and depends on the first guess and other factors. The slowest method is LU QUAD, for which a large amount of computation time is devoted to the translation of double-precision real to quadruple-precision real and back. Figure 5 also shows the computing time required to obtain the reference solution, which requires more time.

The UMF method is the slowest double-precision direct method, but its multifrontal block strategy becomes interesting for large systems. The resolution of the Morel-Morgan test requires 33 times more CPU time than the resolution of the MoMaS easy test for the UMF method, whereas it takes 190 times more time for the LU method.

Among the iterative methods, the fastest is the Gauss-Seidel method and the slowest is the Inc. CHOLESKY method. The two most robust iterative methods, BiCGStab and GMRES, are rapid, sometimes more so than the direct robust methods, LU and UMF, especially for large systems (Morel-Morgan test case). GMRES is less case-dependent than BiCGStab, leading to similar computing time, regardless of the test case.

As expected, introducing preconditioning techniques (LU Equil, DGESVX, and GMRES Equil) or decreasing the convergence criteria for an iterative method (GMRES 10^{-12}) leads to increased computing time. The computing time for preconditioning does not depend only on the system's size: the Valocchi, MoMaS easy, and MoMaS medium test cases (system size of 5 ×5) are solved with the same computing time for all the direct methods, but their resolution when using LU QUAD Equil, LU Equil, and GMRES Equil is faster.

Appendix 6 shows the computation time (log scale) for each test case and each method depending on the condition number. Appendix 6 (F-1 to F-10) shows that, as expected, the computation time of the direct methods does not depend on the condition number of the system. The LU method is usually 10 times faster than the UMF method, except for the Morel-Morgan test case, in which LU is only 1.5 times faster.

In Appendix 6 (F-11 to F-20), the general tendency for the iterative methods is to require the same computation time, independent of the condition number. The oscillations presented by the curves seem to be not related to the condition number. For the test case without minerals, the Gauss-Seidel method is efficient. The two most robust methods, BiCGStab and GMRES, are often the third and

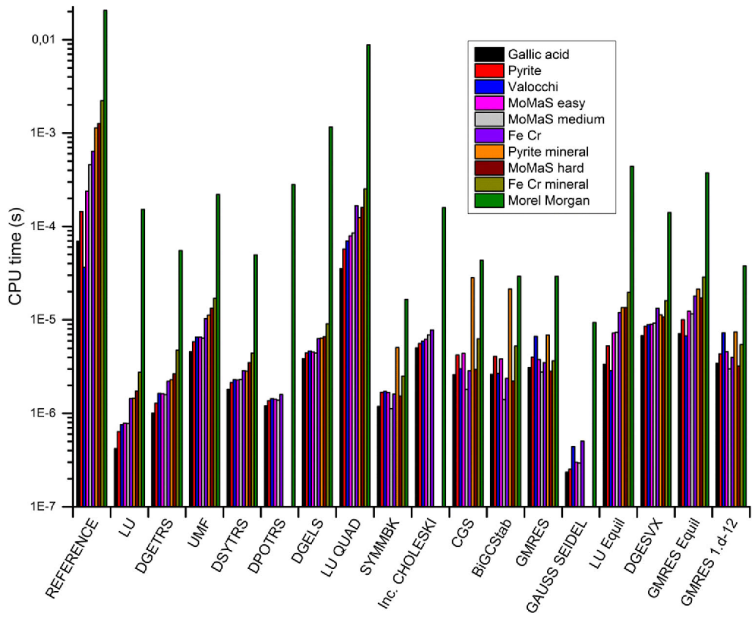


Fig. 5 CPU time (s) for each method and each test case

fourth fastest methods (Gauss-Seidel and SYMMBK are the fastest).

4 Proposal of a new algorithm

Based on our results, we propose an algorithm to optimize the resolution of a chemical system using a Newton-Raphson-like method.

Examining the failure ratio results, seven methods are eligible: LU and LU QUAD as direct methods, GMRES and Gauss-Seidel (if no minerals) as iterative methods, LU Equil and GMRES Equil as preconditioned methods, and the reference method (LU QUAD Equil).

Because these methods are included in a Newton minimization procedure, the most important accuracy criterion is the direction of the minimization, i.e. the angle between the reference and the calculated solution. The behaviour of this direction is strongly correlated with the condition number of the system and is correct if the condition number is less than the critical value and wrong if the condition number is greater than the critical value (see Appendix 5). The critical condition number is 10^8 for GMRES, 10^{16} for the double-precision direct methods, 10^{32} for LU QUAD, and case-dependent for preconditioned methods (10^{20} to 10^{60}). Gauss-Seidel leads to wrong directions for very low condition numbers (Appendix 5 (E-11 and E-12)).

In terms of efficiency, the most rapid method is Gauss-Seidel when it is available. The second most efficient method is LU for small systems (less than 10×10) or GMRES for larger systems (more than 10×10), and the slowest method is LU QUAD. For small systems (less than 5×5), LU Equil is as fast as GMRES but becomes slower as the system size increases.

Table 4 Algorithm for equilibrium computation

cond(Z)	Inversion method
$>10^{30}$	LU QUAD Equil
$10^{30} \geq \text{cond}(Z) > 10^{14}$	LU QUAD
$10^{14} \geq \text{cond}(Z) > 10^{4}$	LU (Nx + NcP < 10) GMRES (Nx + NcP ≥ 10)
$10^{4} > \text{cond}(Z)$	LU

We recommend using LU, LU QUAD, GMRES, and the reference method LU QUAD Equil. Gauss-Seidel should be rejected because of its wrong direction, and equilibration does not sufficiently improve the behaviour of double-precision routines.

Using Eq. 25, it is possible to estimate the condition number of the system without additional computation. This estimation enables the selection of the best-adapted method depending on the system size and condition number.

The goal is to use the most robust method (LU QUAD with preconditioning) for high condition number systems (more than 10^{32}) in the first Newton-Raphson iterations. When the condition number is sufficiently decreased, the preconditioning becomes useless and LU QUAD can be used until the condition number is less than 10^{16}. Then, a faster method is used to obtain a coarse approximation of the solution, LU for small systems and GMRES for large systems (more than 10×10). To find the exact solution, the LU direct method is used.

We propose the algorithm presented in Table 4 and compare it with several inversion methods in a Newton-Raphson algorithm. The 10 chemical test cases are solved

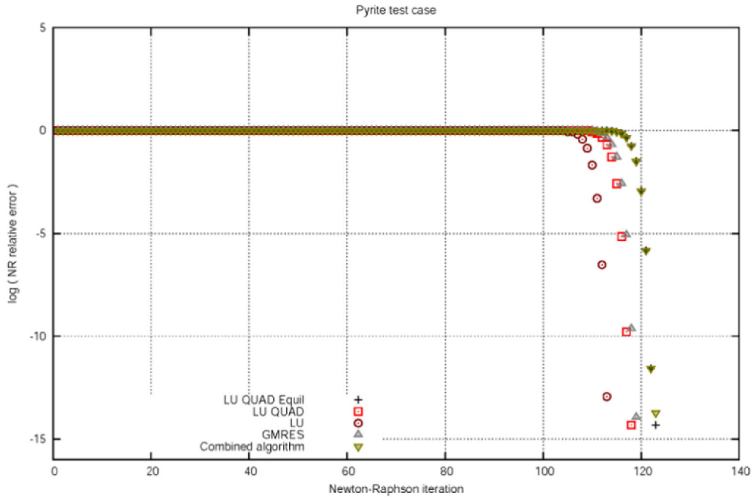


Fig. 6 Evolution of $NR_{relativeerror}$ as a function of the Newton-Raphson iteration for the pyrite test case

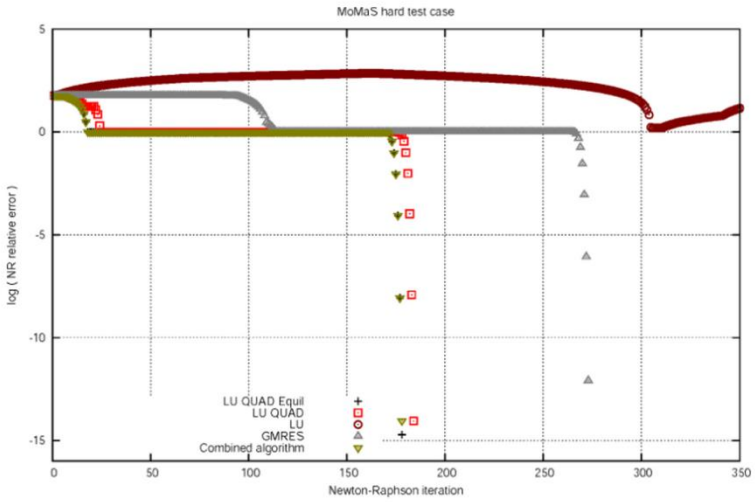


Fig. 7 Evolution of $NR_{relativeerror}$ as a function of the Newton-Raphson iteration for the MoMaS hard case

using the combined algorithm or one of the selected methods: LU QUAD Equil (used as the reference solution), LU QUAD, LU, and GMRES. Appendix 7 shows the evolution of the $NR_{relativeerror}$ (7) as a function of the Newton-Raphson iterations

Figure 6 shows that all the methods are equivalent for easy test cases (see Appendix 7 (G-1 to G-3)). Nevertheless, the use of LU inversion leads to non-convergence, even if the test is easy, as observed for the Valocchi test (Appendix 7 (G-2)). If the difficulty of the test increases, the lower

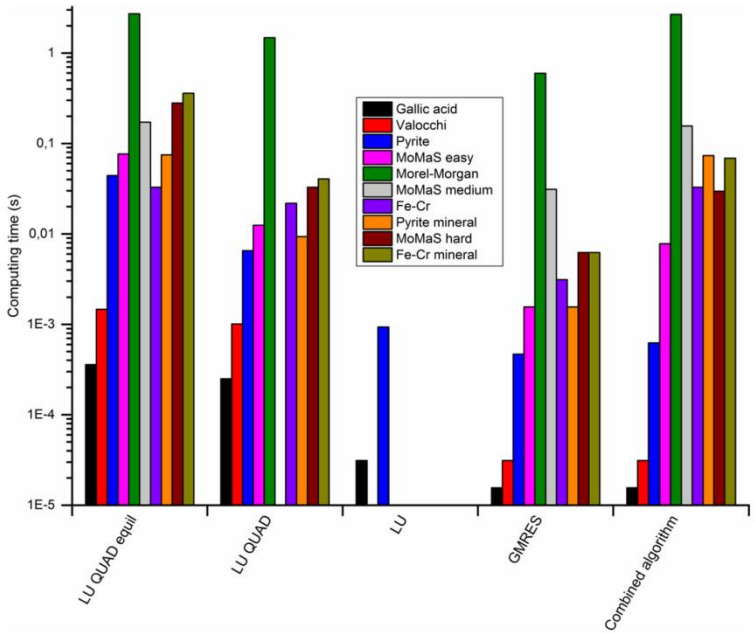


Fig. 8 Computation time (s) as a function of test case and algorithm

accuracy of GMRES (compared to the quadruple-precision routine used in LU QUAD Equil, LU QUAD, and the combined algorithm) leads to a greater number of Newton iterations, as shown in Fig. 7 for the MoMaS hard case. This point is confirmed for other cases (see Appendix 7 (G-4 to G-9)). For the Fe-Cr mineral case (see Appendix 7 (G-10)), only LU QUAD Equil and the combined algorithm can solve the problem. Other methods lead to non-convergence, due to overflow for the GMRES algorithm (overflow appears in the Newton algorithm and is not due to GMRES itself) and because LU QUAD and LU are unable to give an accurate descent direction.

Appendix 7 (G-11) shows the evolution of the relation between the norm of Y and the condition number of the Jacobian matrix during the minimization process. This figure is similar to Fig. 1, confirming the empirical relation (25). This relation cannot be used close to the solution, and the condition number tends to be a case-dependent limit for very low $||Y||$.

Nevertheless, the number of iterations is not the critical point. Because the time required by one iteration changes depending on the method used, we have to consider the total computation time. By plotting the total computation time required to solve each test case depending on the algorithm used (see Fig. 8), we can see that

(i) LU QUAD Equil, as expected, is the slowest. Nevertheless, this method allows the convergence of the Newton-Raphson method for all test cases.
(ii) LU QUAD is slightly faster. The difference between LU QUAD Equil and LU QUAD gives an indication of the time used for matrix equilibration. This time is greater for pyrite, MoMaS easy, pyrite mineral, MoMaS hard, and Fe-Cr mineral than for the other test cases.
(iii) LU is fast when it leads to convergence, but this method results in a very weak Newton-Raphson algorithm.
(iv) GMRES always results in the fastest Newton-Raphson algorithm. It has been shown (Fig. 7, Appendix 7 (G-8)) that the number of required iterations can be twice the number for other methods, but we show (Fig. 5) that the GMRES method is faster than the other methods.
(v) The proposed combined algorithm leads to intermediate computing times, equivalent to those of LU QUAD Equil and LU QUAD, depending on the case.

According to our results, GMRES should be systematically used because it is fast and usually leads to convergence of the Newton-Raphson algorithm. The combined algorithm should be used for very high condition numbers or for recomputing a failed run.

5 Conclusion

In this work, we focus on the resolution of small linear systems generated using the Newton-Raphson algorithm to solve equilibrium chemistry problems. For the first time, we propose a study of the condition number of such linear systems and find that the range of values covered is unusually large. This characteristic leads to specific numerical problems, with matrices that are quite small (approximately 10×10) but very badly conditioned (up to 10^{100}). Ten different chemical systems are studied.

There is a strong linear relationship between the logarithm of the condition number of the matrix and the logarithm of the norm of the objective function. This factor can be exploited to create efficient algorithms. This relation is strictly an empirical one and is not valuable for low condition numbers.

A wide variety of linear solvers have been tested, and several direct and iterative solvers are selected. Some of these solvers are specific for a class of matrix, symmetric or positive definite, while others are generic. A preconditioning method (matrix equilibration) has also been tested to reduce the conditioning of the systems.

According to our selected test cases, only the LU and LU QUAD direct methods, the GMRES iterative method, and LU Equil and GMRES Equil preconditioned methods are sufficiently robust to solve all the tests.

According to the size of the chemical tests, the LU method is faster than the GMRES method. However, our results for the Fe-Cr mineral and Morel-Morgan cases show that GMRES is preferable for larger chemical systems (more than 10 components). Chemical systems with more than 10 components have not been frequently modelled in the past decade. However, the use of geochemical databases makes the construction of large geochemical systems easier, and the increase in computation capacities makes it possible. For very large geochemical systems, we recommend the GMRES method.

We also propose using the linear relationship between the condition number of the Jacobian matrix and the norm of the objective function to develop an efficient algorithm.

The classic LU method is not a good choice. Its weakness is its low robustness for challenging test cases. We recommend using the GMRES method, which is fast and usually leads to convergence of the Newton-Raphson algorithm. For very high condition numbers (more than 10^{100}), we recommend the most robust LU QUAD Equil method. When the Newton-Raphson method is sufficiently near the solution to decrease the condition number, the faster GMRES method can be used. By using the linear relationship between $\mathrm{cond}(Z)$ and $||Y||$, the transition between the two methods can be achieved without computing the condition number (which is very expensive).

This work explores a new research field by studying geochemical computation from a condition number point of view. We attempted to benchmark a wide variety of linear solvers, but it was not possible to explore the flexibility of all the tested solvers. This study will help us to eliminate some solvers so that our future work can focus on the most promising: LU, LU QUAD, GMRES, LU Equil, and GMRES Equil. Some points for future exploration are as follows:

(i) We did not extensively test the robustness and the efficiency of the Newton-Raphson algorithm. Further work should examine the influence of the initial Newton-Raphson guess to confirm our conclusions about the high efficiency of the GMRES method.
(ii) The accuracy of iterative methods depends on the value of the convergence criterion (which we set to 10^{-8}) and on the method used to check the convergence (we used the default method). Moreover, the efficiency can vary depending on the initial guess provided by the user. In this work, we used the easiest initial guess: the residual for the tests from the Newton-Raphson method and the previous Newton-Raphson step for the test in a Newton-Raphson algorithm. We believe that it is possible to make a better choice, markedly enhancing the efficiency of the iterative methods.
(iii) The GMRES method allows the use of left and/or right preconditioners. These preconditioners can increase the robustness, accuracy, and efficiency of the method. More generally, several classes of preconditioners that may reduce the condition number of the linear system can be used [65, 66]. In this work, we explored the use of one preconditioner: matrix equilibration. However, other classes of preconditioners may be more efficient.
(iv) Previous works have addressed the use of methods to solve geochemical equilibria other than the Newton-Raphson method [17, 44, 49, 67]. It has been shown [17] that an efficient algorithm can be obtained by combining a zero-order method with the Newton-Raphson approach.
(v) The size of the chemical tests presented here is representative of the sizes actually used in environmental studies. We have shown that the GMRES method may be efficient for large systems. In anticipation of future needs, it may be useful to test chemical systems larger than the Morel-Morgan system.
(vi) Part of the Newton minimization related to very large condition numbers (far from the solution) can be performed using *random* methods; GMRES is efficient even though its descent direction is not accurate for high condition numbers. Some methods, such as simulated annealing and particle swarm optimization, could be used in future research.

These factors should be explored in light of the results presented in this study. We proposed a large set of chemical tests, a criterion to determine the difficulty of these tests (the condition number), and a panel of numerical methods that should be studied preferentially.

As a more general consideration, the reader should pay particular attention to the old Morel-Morgan test case and the more realistic pyrite test case. The Morel-Morgan test uses Fe^{2+} and Fe^{3+}, Co^{2+} and Co^{3+}, and SO_4^{2-} and S^{2-} as components whereas the pyrite case uses O_2, Fe^{2+}, and SO_4^{2-}. The first studies on geochemical computation avoided redox problems. We show that redox problems lead to higher condition numbers because the stoichiometric coefficients and equilibrium constants cover a wider range. Several geochemical databases avoid the introduction of redox reactions. There is sometimes a good reason to not write redox reactions as equilibria (slow reaction rates, irreversible reactions) as done in Arora et al. [2]. However, the reason is sometimes numeric, and redox reactions are avoided because they lead to non-convergence.

We propose the use of quadruple-precision real for challenging chemical systems. In this work, the core of the geochemical code is conserved as double-precision real, and only the linear system tool is set as quadruple precision. Rewriting an entire geochemical code in a quadruple-precision format will result in robust code but at the cost of an important and rebarbative work as well as efficiency. In this stage of our research, we do not recommend such an effort because implementing LU decomposition using quadruple-precision real is very efficient, requiring only a minor modification of existing code and reducing the computation time.

Acknowledgments Hela Machat has been supported by a grant from the Tunisian Government. This work is supported by the BRGM-CNRS CUBICM project. We thank the reviewers for their helpful comments.

Compliance with Ethical Standards Compliance with ethical standards

Conflict of interests The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Walter, A.L. et al.: Modeling of multicomponent reactive transport in groundwater. 2. Metal mobility in aquifers impacted by acidic mine tailings discharge. Water Resour. Res. **30**(11), 3149–3158 (1994)
2. Arora, B. et al.: A reactive transport benchmark on heavy metal cycling in lake sediments Computational Geosciences (2014)
3. De Windt, L., Leclercq, S., Van der Lee, J.: Assessing the durability of nuclear glass with respect to silica controlling processes in a clayey underground disposal. In: 29th International Symposium on the Scientific Basis for Nuclear Waste Management XXIX. Materials Research Society Symposium Proceedings, Ghent; Belgium (2005)
4. Hoteit, H., Ackerer, P., Mose, R.: Nuclear waste disposal simulations: Couplex test cases. Comput. Geosci. **8**(2), 99–124 (2004)
5. Tompson, A.F.B., et al.: On the evaluation of groundwater contamination from underground nuclear tests. Environ. Geol. **42**(2-3), 235–247 (2002)
6. Andre, L., et al.: Numerical modeling of fluid-rock chemical interactions at the supercritical CO2-liquid interface during CO2 injection into a carbonate reservoir, the Dogger aquifer (Paris Basin, France). Energy Convers. Manag. **48**(6), 1782–1797 (2007)
7. Kang, Q., et al.: Pore scale modeling of reactive transport involved in geologic CO2 sequestration. Transp. Porous Media **82**(1), 197–213 (2010)
8. Navarre-Sitchler, A.K., et al.: Elucidating geochemical response of shallow heterogeneous aquifers to CO2 leakage using high-performance computing: implications for monitoring of CO2 sequestration. Adv. Water Resour. **53**(0), 45–55 (2013)
9. Pruess, K. et al.: Code intercomparison builds confidence in numerical simulation models for geologic disposal of CO2. Energy **29**(9-10), 1431–1444 (2004)
10. Regnault, O., et al.: Etude experimentale de la reactivite du CO2 supercritique vis-a-vis de phases minerales pures. Implications pour la sequestration geologique de CO2. Compt. Rendus Geosci. **337**(15), 1331–1339 (2005)
11. Valocchi, A.J., Street, R.L., Roberts, P.V.: Transport of ion-exchanging solutes in groundwater: chromatographic theory and field simulation. Water Resour. Res. **17**, 1517–1527 (1981)
12. Lichtner, P.C.: Continuum model for simultaneous chemical reactions and mass transport in hydrothermal systems. Geochim. Cosmochim. Acta **49**(3), 779–800 (1985)
13. Appelo, C.A.J.: Hydrogeochemical transport modelling. Proceed. Inf.—Comm. Hydrol. Res. TNO **43**, 81–104 (1990)
14. Yeh, G.T., Tripathi, V.S.: A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components. Water Resour. Res. **25**, 93–108 (1989)
15. Carrayrou, J. et al.: Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems—the MoMaS benchmark case. Computational Geosciences **14**(3), 483–502 (2010)
16. Hammond, G.E., Valocchi, A.J., Lichtner, P.C.: Modeling multicomponent reactive transport on parallel computers using Jacobian-Free Newton Krylov with operator-split preconditioning. In: Hassanizadeh, S.M. (ed.) Developments in water science, computational methods in water resources, Proceedings of the XIVth International Conference on Computational Methods in Water Resources (CMWR XIV), pp. 727–734. Elsevier (2002)
17. Carrayrou, J., Mosé, R., Behra, P.: New efficient algorithm for solving thermodynamic chemistry. AIChE J. **48**(4), 894–904 (2002)
18. Amir, L., Kern, M.: A global method for coupling transport with chemistry in heterogeneous porous media. Comput. Geosci. **14**(3), 465–481 (2010)
19. Quarteroni, A., Sacco, R., Saleri, F.: Numerical mathematics. In: Marsden, J.E., Sirovich, L., Antman. S.S. (eds.) Texts in Applied Mathematics. 2nd edn. Springer, Heidelberg (2007)
20. Axelsson, O., et al.: Direct solution and incomplete factorization preconditioned conjugate gradient methods. Comparison of algebraic solution methods on a set of benchmark problems in linear elasticity, in STW report. 2000, Department of Mathematics, Catholic University of Nijmegen: Nijmegen, The Netherlands. pp. 1-36
21. Barrett, R., Berry, M., Chan, T.F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., Van Der Vorst, H. Templates for the solution of linear systems: building blocks for iterative methods, 2nd edn. SIAM, Philadelphia (1994)
22. Gould, N.I.M., Hu, Y., Scott, J.A.: A numerical evaluation of sparse direct solvers for the solution of large sparse, symmetric linear systems of equations. 2005, Council for the Central Laboratory of the Research Councils
23. Allaire, G., Kaber, S.M. In: Marsden, J.E., Sirovich, L., Antman, S.S. (eds.): Numerical linear algebra. Texts in applied mathematics. Springer, New York (2008)
24. Baldwin, C., et al.: Iterative linear solvers in a 2D radiation-hydrodynamics code: methods and performance. J. Comput. Phys. **154**(1), 1–40 (1999)
25. Chao B.T., L.H.L., Scott, E.J.: On the solution of ill-conditioned, simultaneous, linear, algebraic equations by machine computation, in University of Illinois Bulletin. 1961, University of Illinois
26. Hadjidimos, A.: Successive overrelaxation (SOR) and related methods. J. Comput. Appl. Math. **123**(1-2), 177–199 (2000)
27. Kalambi, I.B.: A comparison of three iterative methods for the solution of linear equations. J. Appl. Sci. Environ. Manag. **12**(4), 53–55 (2008)
28. Klisinski, M., Runesson, K.: Improved symmetric and non-symmetric solvers for FE calculations. Adv. Eng. Softw. **18**(1), 41–51 (1993)
29. Schenk, O., Gartner, K.: Solving unsymmetric sparse systems of linear equations with PARDISO. Fut. Gener. Comput. Syst. **20**(3), 475–487 (2004)
30. Xue, X.J. et al.: A direct algorithm for solving ill-conditioned linear algebraic systems. JCPDS-Int. Centre Diffract. Data Adv. X-ray Anal. **42**, 629–633 (2000)
31. Pyzara, A., Bylina, B., Bylina, J.: The influence of a matrix condition number on iterative methods' convergence (2011)
32. Hoffmann, J., Kräsutle, S., Knabner, P.: A parallel global-implicit 2-D solver for reactive transport problems in porous media based on a reduction scheme and its application to the MoMaS benchmark problem. Comput. Geosci. **14**(3), 421–433 (2010)
33. Soleymani, F.: A new method for solving ill-conditioned linear systems. Opuscula Math. **33**(2), 337–344 (2013)
34. Morel, F., Morgan, J.: A numerical method for computing equilibria in aqueous chemical systems. Environ. Sci. Technol. **6**(1), 58–67 (1972)
35. Morel, F.M.M.: Principles of aquatic chemistry. Wiley Interscience, New York (1983)
36. De Windt, L. et al.: Intercomparison of reactive transport models applied to UO2 oxidative dissolution and uranium migration. J. Contam. Hydrol. **61**(1-4), 303–312 (2003)
37. Jauzein, M. et al.: A flexible computer code for modelling transport in porous media: impact. Geoderma **44**(2–3), 95–113 (1989)
38. Parkhurst, D.L., Appelo, C.A.J.: User's guide to PHREEQC (version 2)—a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. Water Resour. Invest., Editor. 1999: Denver. p. 312
39. Van der Lee, J.: CHESS another speciation and surface complexation computer code. E.d.M.d. Paris, Editor. 1993: Fontainebleau. p. 85

40. Westall, J.C.: MICROQL: a chemical equilibrium program in BASIC. Computation of adsorption equilibria in BASIC. S.F.I.o.T. EAWAG, Editor. 1979: Dübandorf. p. 42
41. Westall, J.C.: FITEQL ver. 2.1. 1982: Corvallis
42. Westall, J.C., Zachary, J.L., Morel, F.M.M.: MINEQL: a computer program for the calculation of chemical equilibrium composition of aqueous system. R.M.P. Laboratory, Editor. 1976: Cambridge. p. 91
43. Walter, L.J., Wolery, T.J.: A monotone-sequences algorithm and FORTRAN IV program for calculation of equilibrium distributions of chemical species. Comput. Geosci. **1**, 57–63 (1975)
44. Wigley, T.M.L.: WATSPEC: a computer program for determining the equilibrium speciation of aqueous solutions. B.G.R.G. Tech. Bull., Editor. 1977. p. 49
45. Jennings, A.A., Kirkner, D.J., Theis, T.L.: Multicomponent equilibrium chemistry in groundwater quality models. Water Resour. Res. **18**, 1089–1096 (1982)
46. Cederberg, A., Street, R.L., Leckie, J.O.: A groundwater mass transport and equilibrium chemistry model for multicomponent systems. Water Resour Res. **21**, 1095–1104 (1985)
47. Yeh, G.T., Tripathi, V.S.: A model for simulating transport of reactive multispecies components: model development and demonstration. Water Resour. Res. **27**(12), 3075–3094 (1991)
48. Carrayrou, J.: Looking for some reference solutions for the reactive transport benchmark of MoMaS with SPECY. Comput. Geosci. **14**(3), 393–403 (2010)
49. Brassard, P., Bodurtha, P.: A feasible set for chemical speciation problems. Comput. Geosci. **26**(3), 277–291 (2000)
50. Carrayrou, J., Kern, M., Knabner, P.: Reactive transport benchmark of MoMaS. Comput. Geosci. **14**(3), 385–392 (2010)
51. Fendorf, S.E., Li, G.: Kinetics of chromate reduction by ferrous iron. Environ. Sci. Technol. **30**(5), 1614–1617 (1996)
52. Chilakapati, A. et al.: Groundwater flow, multicomponent transport and biogeochemistry: development and application of a coupled process model. J. Contam. Hydrol. **43**(3-4), 303–325 (2000)
53. Knight, P., Ruiz, D., Ucar, B.: A symmetry preserving algorithm for matrix scaling. SIAM J. Matrix Anal. Appl. **35**(3), 931–955 (2014)
54. Golub, H.V., Van Loan, C.F.: Matrix computations. 3rd ed. The Johns Hopkins University Press, Baltimore (1996)
55. Davis, T.A., Duff, I.S.: A combined unifrontal/multifrontal method for unsymmetric sparse matrices. ACM Trans. Math. Softw. **25**(1), 1–20 (1999)
56. Woźnicki, Z.: On performance of SOR method for solving non-symmetric linear systems. J. Comput. Appl. Math. **137**(1), 145–176 (2001)
57. Saad, Y., Van Der Vorst, H.A.: Iterative solution of linear systems in the 20th century. J. Comput. Appl. Math. **123**(1-2), 1–33 (2000)
58. Diersch, H.J.G.: FEFLOW reference manual. DHI-WASY GmbH, Berlin (2009)
59. Van der Lee, J., et al.: Presentation and application of the reactive transport code HYTEC. In: Hassanizadeh, S.M. (ed.) Developments in Water Science, Computational Methods in Water Resources, Proceedings of the XIVth International Conference on Computational Methods in Water Resources (CMWR XIV), pp. 599–606. Elsevier (2002)
60. Press, W.H., S.A.T., Vettering, W.T., Flannery, B.P. Numerical recipes in FORTRAN: the art of scientific computation, 2nd edn., pp. 123–124. Cambridge University Press, New Yor (1992)
61. The Linear Algebra Package (LAPACK) can be obtained free of charge from the address listed here: http://www.netlib.org/lapack
62. Kincaid, D., Cheney, W. Numerical analysis: mathematics of scientific computing, 3rd edn. American Mathematical Society (2002)
63. HSL: A collection of Fortran codes for large scale scientific computation. http://www.hsl.rl.ac.uk (2013)
64. Chapter 8 Systems of nonlinear equations. In: Studies in computational mathematics, Claude, B. Editor. 1997, Elsevier. pp. 287–336
65. Soleymani, F.: A rapid numerical algorithm to compute matrix inversion. Int. J. Math. Math. Sci. **2012** (2012)
66. Soleymani, F.: On a fast iterative method for approximate inverse of matrices. Commun. Korean Math. Soc. **28**(2), 407–418 (2013)
67. Morin, K.A.: Simplified explanations and examples of computerized methods for calculating chemical equilibrium in water. Comput. Geosci. **11**, 409–416 (1985)

Annexe 5. Thermodynamic equilibrium solutions through a modified Newton-Raphson method

AIChE

Thermodynamic Equilibrium Solutions Through a Modified Newton Raphson Method

Marianna Marinoni, Jérôme Carrayrou, Yann Lucas, and Philippe Ackerer
LHyGeS, Université de Strasbourg/EOST-CNRS, 1, rue Blessig, 67000 Strasbourg, France

DOI 10.1002/aic.15506
Published online in Wiley Online Library (wileyonlinelibrary.com)

In numerical codes for reactive transport modeling, systems of nonlinear chemical equations are often solved through the Newton Raphson method (NR). NR is an iterative procedure that results in a sequential solution of linear systems. The algorithm is known for its effectiveness in the vicinity of the solution but also for its lack of robustness otherwise. Therefore, inaccurate initial conditions can lead to non-convergence or excessive numbers of iterations, which significantly increase the computational cost. In this work, we show that inaccurate initial conditions can lead to very ill-conditioned system matrices, which makes NR inefficient. This efficiency is improved by preconditioning techniques and/or by coupling the NR method with a zero-order method called the positive continuous fraction method. Numerical experiments that are based on seven different test cases show that the ill-conditioned linear systems within NR represent a problem and that coupling NR with a method that bypasses the computation of the Jacobian matrix significantly improves the robustness and efficiency of the algorithm. © 2016 American Institute of Chemical Engineers *AIChE J*, 00: 000–000, 2016

Keywords: computational chemistry, environmental engineering

Introduction

Reactive transport modeling is applied in different fields of science and engineering, including combustion, catalysis, atmospheric chemistry, water chemistry, and geochemistry. Reactive transport modeling copes with the solution of transport equations that are coupled with biogeochemical reactions. In this context, two approaches exist: a *global implicit* (or *one step*) approach, and a *sequential iterative* or *sequential non-iterative* approach (the so-called operator splitting approach).[1 3] The global implicit approach consists of introducing reaction equations into transport equations and solving the resultant system, while the operator splitting approach consists of sequentially solving transport equations and biogeochemical reactions.

Although the numerical results that are provided in this article can be widely applied to all reactive transport simulations, we use reactive transport in soils and groundwater resources for illustration in this study. Several numerical codes are available to simulate reactive transport in this type of porous material.[3] Some of these codes adopt the global implicit approach, such as PFLOTRAN[4] or MIN3P[5,6]; others adopt the operator splitting approach, such as HPx,[7 9] PHT3D,[10] HYTEC,[11] TOUGHREACT,[12 15] or eSTOMP[16,17]; and others allow the user to choose between the two (e.g., Crunchflow,[18] HYDROGEOCHEM[19,20]) Each methodology has its own advantages and disadvantages, as determined by research that has been conducted on these methods.[1,2] The operator-splitting approach is often performed by coupling modules that solve transport equations with modules that are designed to solve biogeochemical equations. For instance, HPx and PHT3D rely on the geochemical code PHREEQC[3,21] to solve for the chemistry, while the geochemical portion of HYTEC[11] is solved by the code CHESS.[3,22,23] The platform OpenGeoSys[24] has been specifically developed to facilitate interactions between modules that deal with problems from different fields. Examples of geochemical modules that are designed to be coupled with transport codes include CHEPROO[25] and MINTEQA2.[26] In the context of the operator splitting approach, the chemical equations must be solved potentially several thousands of times per time step, once for each cell/node[22] of the mesh that is designed to solve the transport equation (typically, several tens of thousands of cells/nodes for 2D problems and several hundreds of thousands for 3D). Moreover, the entire transport computation must be repeated with a smaller time step when the chemical system cannot be resolved. Thus, efficient (i.e., robust and fast) solution techniques are mandatory. In the majority of geochemical codes, a nonlinear system of equations is solved through Newton Raphson (NR)-based algorithms. Less popular methods exist like simplex method[27] for equilibrium problems.

The NR method is an iterative procedure that provides a solution of nonlinear systems of chemical equations through the repeated resolution of linear systems. As with all iterative procedures, NR requires a set of initial guesses to begin its path to the solution. This method is appreciated for its quadratic rate of convergence when favorable initial guesses are picked; conversely, an unlucky combination of initial guesses will likely prevent the algorithm from converging.[22] For this reason, the NR method is often implemented alongside techniques that utilize a preliminary selection of initial solutions.[21]

Correspondence concerning this article should be addressed to P. Ackerer at ackerer@unistra.fr

When previous time solutions are available, they represent good initial guesses for the next iterations. This is the case for the simulation of a close system for example. For reactive transport simulation, initial conditions are not always available since solutes are transported from one cell/element of the discretized domain to the neighbor cells/elements. This may introduce significant changes in concentrations especially for advective dominant transport and the initial conditions can be far from the solution.

Depending on the problem, the coefficients that appear in a system of equations can be very different, potentially causing linear systems to be ill-conditioned.[28] The accuracy of the numerical solution of an ill-conditioned system can be very poor and may provide results that are very different from the exact solution, i.e., the wrong descent direction of an iterative procedure. Techniques such as *line search*[29] (or *one dimensional search*[28]) or similar methods[22] that deal with the amplitude of the step but leave the direction unchanged are exposed to the same risk. The motivation of this article is to explore the effects of ill-conditioned linear systems from NR on the outcome of the algorithm and to compare different solutions to improve its efficiency. Several techniques exist to ameliorate the condition of a linear system like precondioning[30] or to improve the solution accuracy like QR factorization.[31] In the context of reactive transport, computations are likely to be called thousands of times, so we studied the impact of the simplest preconditioning technique, which is known as *scaling*,[32] on the evaluation of thermodynamic equilibrium for a selection of numerically challenging problems. A declination of scaling techniques is proposed and their consequences on the robustness of the NR method are presented.

In practice, the NR method is often coupled with other algorithms, so we compared the effects of scaling techniques with the effects of coupling the NR algorithm with the positive continuous fraction method,[33] an effective zero-order technique that was initially proposed to treat linear concentrations. Here, we adapt this method to work with concentrations on a logarithmic scale.

Thermodynamic Equilibrium: Governing Equations

Modeling chemical reactions at equilibrium is a basic feature in many reactive transport codes and is the principal objective of other computational modules. Although not a unique alternative,[34,35] thermodynamic equilibrium is often described through a combination of mass conservation and mass action laws that are written in terms of *species* and *components*.[2] The formulation of species and components has been adopted by many modelers (Reed,[36] Westall,[37] Cedeberg,[38] Yeh and Tripathi,[39] Steefel and Lasaga,[40] Parkhurst and Appelo[21]) and is currently implemented in codes such as CHESS,[23] CHEPROO,[25] TOUGHREACT,[15] Crunchflow,[18] PHREEQC[21] among others.

Chemical reactions at equilibrium can also be considered as kinetic since their rate, although sometimes high, is always finite (Steefel,[41] Chilakapati,[42,43] Fang[44]). The whole reaction network is then represented through a system of ordinary differential equations to be integrated in time. When great differences between kinetic rates arise, the system to be solved becomes stiff. Its solution is then challenging and might require many time steps to reach steady state (equilibrium). Of course, if the number of time steps required to solve the problem needs more computer time than an iterative procedure used to solve equilibrium, the kinetic approach is not efficient. Therefore, when the kinetic system becomes too difficult to solve, the modeling of the fastest reactions through thermodynamic equilibrium is highly recommended (Steefel,[41] Chilakapati,[43] Fang,[44] Leal[45])

One can individuate a subset of *components* within all *chemical species* in a chemical system at thermodynamic equilibrium to entirely describe the system. Components (often addressed as primary species) are linearly independent, and their combinations recreate all chemical species (secondary species). The relationship between components and other chemical species is mathematically expressed as follows

$$\sum_{j=1}^{Nx} b_{i,j} X_j \Longleftrightarrow C_i \quad i=1,\ldots,N_C\,, \tag{1}$$

where X_j represents a generic component; C_i is a generic chemical species; $b_{i,j}$ is a generic stoichiometric coefficient; and N_X and N_C are the number of components and dissolved chemical species, respectively. Equation 1 describes a qualitative relationship between components and other chemical species and does not provide quantitative information regarding the concentrations or activities of different elements. Quantitative relationships are provided by the conservation law (2) and mass action law (3)

$$\left[T_j\right] = Total(X_j) = \sum_{i=1}^{Nc} b_{i,j}[C_i] \quad j=1,\ldots,N_X \tag{2}$$

$$\{C_i\} = K_i \prod_{j=1}^{Nx} \{X_j\}^{b_{i,j}} \quad i=1,\ldots,N_C \tag{3}$$

where $[-]$ defines a concentration, $\{-\}$ defines an activity, $\left[T_j\right]$ is the total concentration of component X_j that is conserved in the system through chemical reactions (expressed in moles per unit volume), K_i is the thermodynamic equilibrium constant, and $b_{i,j}$ is a generic stoichiometric coefficient that is related to the formation of chemical species C_i based on the components X_j. If present, precipitates that constitute the solid phase are identified with the symbol Cp_l, with $l=1,..,N_{Cp}$, where N_{Cp} is the total number of precipitates. Thus, Eq. 2 undergoes a modification

$$\left[T_j\right] = Total(X_j) = \sum_{i=1}^{Nc} b_{i,j}[C_i] + \sum_{l=1}^{Ncp} bp_{l,j}[Cp_l] \quad j=1,\ldots,N_X\,, \tag{4}$$

where $[Cp_l]$ is the concentration of precipitate l. Cp_l and $bp_{l,j}$ is the related stoichiometric coefficient. The concentrations of precipitates must be computed within the chemical equilibrium solution, but their activity $\{Cp_l\}$ remains constant and equal to one because the quantity of precipitates is no longer available for reactions. Thus, whenever precipitation occurs, $[Cp_l]$ cannot be deduced through the mass action law and must be treated as an additional unknown.33 is included in the system to balance this supplementary unknown.

$$K_S^l = \prod_j \{X_j\}^{bp_{l,j}} \tag{5}$$

Precipitation occurs only when condition (6) is satisfied

$$\prod_j \{X_j\}^{bp_{l,j}} > K_S^l \tag{6}$$

The relationship between the concentration and activity of a generic species C_i is expressed through Eq. 7, where the activity coefficient γ_i is computed based on the ionic force

$$\{C_i\}=\gamma_i[C_i] \tag{7}$$

Substituting the mass action law (3) into mass conservation (2) while considering the relationship between the concentrations and activities (7) provides a set of N_X equations, where the total concentration can be computed as a sole function of the concentration of components $[X_j]$ and precipitated species $[Cp_l]$

$$[T_j]=\sum_{i=1}^{Nc}\frac{b_{i,j}}{\gamma_i}K_i\prod_{k=1}^{Nx}(\gamma_k[X_k])^{b_{i,k}}+\sum_{l=1}^{Ncp}bp_{l,j}[Cp_l] \quad j=1,\ldots,N_x. \tag{8}$$

In the previous equations, the stoichiometric coefficients $b_{i,j}$ and $bp_{l,j}$ and the thermodynamic equilibrium constant K_i are known. Activity coefficients γ are expressed as nonlinear functions of species' concentrations and therefore of components' concentrations $[X_j]$.

Since the total concentration $[\tilde{T}_j]$ remains unchanged due to mass conservation, the solution strategy consists of finding $[X_j]$ so that $[T_j]=[\tilde{T}_j]$. Therefore, we define the residual Y_j as

$$Y_j=[\tilde{T}_j]-[T_j]=[\tilde{T}_j]-\sum_{i=1}^{Nc}\frac{b_{i,j}}{\gamma_i}K_i\prod_{k=1}^{Nx}(\gamma_k[X_k])^{b_{i,k}}-\sum_{l=1}^{Ncp}bp_{l,j}[Cp_l] \quad j=1,\ldots,N_x. \tag{9}$$

Equation 9 describes a nonlinear system of N_x equations with $N_t=N_x+N_{Cp}$ unknowns. Supplementary equations are then introduced to equilibrate the system based on Eq. 5

$$Y_{N_X+l}=1-\prod_j\frac{(\gamma_j[X_j])^{bp_{l,j}}}{K_S^l} \quad l=1,\ldots,N_{Cp}. \tag{10}$$

To simplify the notation, the N_t unknowns are grouped into a vector $\mathbf{X}$, and the nonlinear system at thermodynamic equilibrium can be rewritten as follows

$$\mathbf{Y}(\mathbf{X})=0. \tag{11}$$

NR Algorithm

One of the most applied algorithms for solving nonlinear systems such as (11) is the NR method.[22] The NR method is iterative and implies the repetition of a given procedure until the solution is reached, i.e., until a given stopping or convergence criterion is satisfied. At each iteration n, the algorithm converges to the solution by updating the unknowns

$$\mathbf{X}_{n+1}=\mathbf{X}_n+\Delta\mathbf{X}_n \tag{12}$$

The increment $\Delta\mathbf{X}_n$ is computed through the solution of a linear system (13)

$$\mathbf{J}_n\Delta\mathbf{X}_n=-\mathbf{Y}_n \tag{13}$$

where $\mathbf{J}_n$ is the Jacobian matrix of the system (11) and $\mathbf{Y}_n$ is the vector of residuals, both of which are computed with the values $\mathbf{X}_n$. The solution of the linear system (13) is usually performed with a dedicated solver. Among the numerous solvers that are available to accomplish this task, we choose LU decomposition[22] with quadruple precision. Quadruple precision reduces the effects of round-off errors and has been shown as the most robust and efficient solver of numerical experiments, as described in Machat and Carrayrou.[46]

The Jacobian matrix $\mathbf{J}$ is an $N_t\times N_t$ matrix that contains the derivatives of each row of the nonlinear system with respect to each unknown. These derivatives are computed analytically and take the following forms

$$J^n_{j,k}\Big|_{\substack{j=1,N_x\\k=1,N_x}}=\sum_{i=1}^{Nc}b_{i,j}\cdot b_{i,k}\frac{[C_i]^n}{[X_k]^n} \tag{14}$$

$$J^n_{j,k}\Big|_{\substack{j=N_x,N_t\\k=N_x,N_t}}=b_{i,j} \tag{15}$$

Because NR is an iterative procedure, the algorithm is initialized with a set of initial guesses for the unknowns (concentrations of components). The NR algorithm is known to be very effective when the initial guesses are picked in the vicinity of the solution but can likely fail to converge if the initial guesses are far from the solution in the unknowns' space.[28] Thus, the NR method is often coupled with other techniques to estimate these initial guesses.[21]

Condition Numbers of a Linear System

The reliability of the solution of the linear system (13) is measured through the evaluation of the condition number of the Jacobian matrix $\mathbf{J}$.[47] The lower the condition number, the more likely the numerical solution of the linear system is to coincide with the exact solution. According to Golub and van Loan,[32] if $\Delta\mathbf{X}_n$ is the solution that is computed by Gaussian elimination and $\Delta\mathbf{X}_{Exact}$ is the exact solution, one can link the relative error with the condition number of J, $\kappa(\mathbf{J})$, and the relative error in the computation of J, $\varepsilon(\mathbf{J})$

$$\frac{\|\Delta\mathbf{X}_n-\Delta\mathbf{X}_{Exact}\|}{\|\Delta\mathbf{X}_{Exact}\|}\leq\kappa(\mathbf{J})\varepsilon(\mathbf{J}) \tag{16}$$

If we assume that the relative error in the computation of J is on the order of the round-off error, 10^{-d}, and that condition number is approximately 10^q, the solution of the system from a direct solver (i.e., LU decomposition) will have at least $d-q$ correct decimal digits. For double-precision real numbers, d is equal to 16, which means that the solution might have no significant digits for $\kappa(\mathbf{J})>10^{16}$. In our context, the exact solution of system (13) coincides with a step toward the solution. Thus, a bad solution of the linear system may result in a poor choice of the direction in our path to thermodynamic equilibrium and, therefore, significantly increase the number of iterations that are required to reach convergence.

The techniques that are used to evaluate the condition number are diverse. In its general definition, this number corresponds to the product between the norm of a given matrix and the norm of its inverse (17), no matter which norm is chosen.[48] In the presence of symmetric and nonsingular matrices,[30] the condition number may also be computed as the ratio between the highest and lowest modules of the eigenvalues λ of the matrix (18). In this particular case, the condition number obtained through Eq. 18 corresponds to the one computed through (17) choosing norm 2 as norm of the Jacobian matrix and its inverse.

$$\kappa(\mathbf{J})=\|\mathbf{J}\|\|\mathbf{J}^{-1}\| \tag{17}$$

$$\kappa(\mathbf{J})=\frac{|\lambda_{\max}|}{|\lambda_{\min}|}=\|\mathbf{J}\|_2\|\mathbf{J}^{-1}\|_2 \tag{18}$$

A possible choice for the norm in Eq. 17 is the norm 1, $\|\mathbf{J}\|_1$ (the maximum absolute column sum), or the infinity norm $\|\mathbf{J}\|_\infty$ (the maximum absolute row sum):

$$\begin{cases} \|\mathbf{J}\|_1=\max_j\left(\sum_i |J_{i,j}|\right) \\ \|\mathbf{J}\|_\infty=\max_i\left(\sum_j |J_{i,j}|\right) \end{cases} \tag{19}$$

The evaluation of the condition number is subordinated to the inversion of the matrix in one case or to the evaluation of the eigenvalues in the other. Therefore, the standard algorithms for the computation of the condition number encounter difficulties in the presence of very ill-conditioned matrices.

The condition number of a given linear system can be modified in different ways. For instance, one can simply multiply its rows or columns by constant values, producing a potentially infinite series of condition numbers. Thus, Buzzi-Ferraris[49] defined *system conditioning* as a particular condition number of a system in what this author called its standard form. This standard form is obtained by dividing each row of the system by its infinity norm, which is computed along the right-hand side of the system and by dividing each column by its infinity norm.[49] This approach is useful because it provides a univocal index to compare results while searching for different and more effective formulations of the problem instead of attempting to ameliorate the condition of a given problem. In this context, searching for a different formulation of the problem means looking for another set of basic components to describe the chemical system. Because this change does not guarantee a decrease of the condition number, this approach will not be studied here.

Working on a Logarithmic Base

A solution that allows some numerical facilities is a change in the variables from Eq. 9. Instead of working with the general unknown concentration $[X_j]$, one should work with the logarithmic transformation of the corresponding activity

$$\xi_j=\ln\left(\{X_j\}\right) \tag{20}$$

The consequences of this transformation on the system are limited. Equation 3 becomes Eq. 21, while Eqs. 9 and 10 take the form of Eqs. 22 and 23

$$\{C_i\}=\exp\left(\ln(K_i)+\sum_j^{N_X} b_{i,j}\xi_j\right) \tag{21}$$

$$Y_j=[\tilde{T}_j]-\sum_{i=1}^{N_C}\frac{b_{i,j}}{\gamma_i}\exp\left(\ln(K_i)+\sum_k^{N_X} b_{i,j}\xi_k\right) + \sum_{l=1}^{N_{CP}} bp_{l,j}[Cp_l]=0 \quad j=1,\ldots,N_x \tag{22}$$

$$\ln K_S^l=\sum_j^{N_X} bp_{l,j}\xi_j \quad l=1,\ldots,N_{Cp} \tag{23}$$

This transformation has the advantage of ensuring the symmetry of the Jacobian matrix and improving its conditioning. This transformation has already been used and implemented, for example, in EQ3/6.[50] The Jacobian matrix entries become as follows (see Appendix A)

$$J^n_{j,k}\Big|_{\substack{j-1,N_x\\k-1,N_x}}=\sum_{i=1}^{N_C} b_{i,j}\cdot b_{i,k}[C_i]^n \tag{24}$$

$$J^n_{j,k}\Big|_{\substack{j-N_x,N_t\\k-N_x,N_t}}=bp_{i,j}\quad J^n_{j,k}\Big|_{\substack{j-k\\j,k\in(N_C+N_{Cp})}}=0 \tag{25}$$

In fact, the concentration of the component j that was present in Eq. 14, disappears from the denominator in Eq. 24, reducing the possibility of fluctuations over an order of magnitude. Equation 24 could be rewritten in its matrix form (26). This notation for the upper left block highlights the influence of the range of concentrations on the condition number

$$\mathbf{J}=\mathbf{B}^T diag(\mathbf{C})\mathbf{B} \tag{26}$$

Preconditioning

Techniques and procedures that are used to reduce the condition number of a given system fall under the definition of *preconditioning*. An accurate summary of these techniques is available in the literature.[30] Some preconditioning techniques, such as preconditioning through an approximate inverse, require consistent computational effort. If the required computational time for preconditioning is significantly higher than some NR iterations, the algorithm will not reduce the computer time that is required to reach the solution. Therefore, we focus our work on methods that improve matrix conditioning without significant computational costs. Those techniques are known as *scaling*.[30,32] Scaling is a procedure that is operated by multiplying one or more rows and/or columns of the linear system by a constant. The only limitation is that multiplying all the lines by the same constant would produce no effects. The idea behind scaling is to solve system (27) instead of the regular linear system that is solved in the standard NR algorithm

$$\mathbf{D}_1\mathbf{J}_n\mathbf{D}_2\Delta\hat{\mathbf{X}}_n=-\mathbf{D}_1\mathbf{Y}_n \tag{27}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices and $\Delta\hat{\mathbf{X}}_n=\mathbf{D}_2^{-1}\Delta\mathbf{X}_n$ is a linear combination of the increments. The previous system is equivalent to (13) but hopefully better conditioned. The entries of $\mathbf{D}_1$ and $\mathbf{D}_2$ may be chosen with respect to the diagonal values of the original matrix $\mathbf{J}$ [51] or according to a given norm of lines and/or columns.[52] Setting $\mathbf{D}_1=\mathbf{I}$ corresponds to performing scaling only over columns, while setting $\mathbf{D}_2=\mathbf{I}$ corresponds to performing scaling only over rows. A further step beyond scaling is matrix equilibration (MEq),[52,53] which sets the norms of rows and columns of a given matrix to a fixed value and repeats a scaling procedure until the norms meet the requirements. However, Golub[32] warned that reactions to scaling are strongly problem dependent and that its implementation is by no means a guarantee of success in the computation of increments.

Scaling Procedures in This Work

In our work, we tested four different combinations of $\mathbf{D}_1$ and $\mathbf{D}_2$ and one procedure for MEq. We tested both scaling techniques that involve rows and columns and techniques that modify only rows. Simple row-scaling of the Jacobian matrix

and its right-hand term is obtained by imposing $\mathbf{D}_2=\mathbf{I}$ and choosing different values of the constants of $\mathbf{D}_1$.

RI: Row Identity scaling

If each row in the Jacobian matrix is defined as a vector $\mathbf{a}_i$ and each element of its diagonal is $J_{i,i}$, one possible choices of the entries of the diagonal matrix is

$$D_{i,i}=\|\mathbf{a}_i\|_\infty \tag{28}$$

where the infinity norm is $\|\mathbf{a}_i\|_\infty=\max|\mathbf{a}_i|$. We refer to scaling through $\mathbf{D}_2=\mathbf{I}$ and the elements of $\mathbf{D}_1$, as in Eq. 28, with the name Row Identity (RI) scaling.

DI: Diagonal Identity scaling

While matrix $\mathbf{D}_2=\mathbf{I}$, the elements of $\mathbf{D}_1$ are picked as in Eq. 29

$$D_{i,i}=J_{i,i} \tag{29}$$

sDsD: Square Diagonal scaling

Two methods of performing scaling that affect include both rows and columns were also tested. The first method was already proposed by Marquardt[51] and is also known by the name *Jacobi preconditioning*.[54] Matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ become

$$\begin{aligned} D_{1,i}&=\sqrt{J_{i,i}}\\ D_{2,i}&=\sqrt{J_{i,i}} \end{aligned} \tag{30}$$

Equation 30 is valid for $i=1,\ldots,Nx$ because the diagonal entries of the matrix become zero (Appendix A) in the presence of precipitates. We refer to this scaling technique with the expression Square Diagonal scaling (sDsD).

RC: Row and Column scaling

The second procedure that acts on both rows and columns (Row and Column scaling—RC) performs only one iteration of a MEq technique that was proposed by Knight, Ruiz, and Uçar.[52] After defining $\mathbf{c}_i$ column-vectors of the matrix $\mathbf{J}$, these authors proposed to choose entries of $\mathbf{D}_1$ and $\mathbf{D}_2$ as in Eq. 31

$$\begin{aligned} D_{1i,i}&=\sqrt{\|\mathbf{a}_i\|_\infty}\\ D_{2i,i}&=\sqrt{\|\mathbf{c}_i\|_\infty} \end{aligned} \tag{31}$$

where $\|\mathbf{a}_i\|_\infty=\max|\mathbf{a}_i|$. If the original matrix was symmetric, then $\mathbf{a}_i=\mathbf{c}_i$, which ensures the symmetry of the scaled matrix.

MEq: Matrix Equilibration

This preconditioning technique consists of repeating RC scaling (Eq. 31) until each RC of the Jacobian matrix has an infinity norm equal to one.

Some remarks on the choice of the preconditioners $\mathbf{D}_1$ and $\mathbf{D}_2$ are necessary. One can also choose $\mathbf{D}_1=\mathbf{I}$, i.e., scaling only the columns of the Jacobian matrix. We avoid this possibility because the accuracy of the system's solution also depends on the right-hand side. Conversely, scaling only the rows of the system prevents us from re-scaling computed increments, which makes RI scaling and Diagonal Identity (DI) scaling the easiest alternatives. Additionally, the scaling of each line or column is not necessary.[55] However, "automatizing" the choice of constant values in some fashion is necessary in the context of an iterative procedure where the solution of the linear system occurs repeatedly. Row Column scaling (RC) is interesting because it can maintain the symmetry of the matrix and reduce gaps on the scale of orders of magnitude between columns, not only between rows, while the effectiveness of MEq has already been proven through the methodology that was proposed by Knight.[56]

PCF Method

The PCF method was presented by Carrayrou[33] as a development of the continuous fraction method, a zero-order method whose variant has been implemented in PHREEQC.[21] According to the PCF method, an approximation of equilibrium can be obtained for a given dissolved component X_j through the iteration of the following equation:

$$[X_j]^{n+1}=\theta_j^n[X_j]^n\left(\frac{sum_j^{prod,n}}{sum_j^{react,n}}\right)^{1/a_{i0,j}}+\left(1-\theta_j^n\right)[X_j]^n \tag{32}$$

where *reactive sum* and *product sum* are defined in Eqs. 33 and 35 and in Eqs. 34 and 36, respectively; θ_j is a weighting factor; and $a_{i0,j}$ is the smallest value of strictly positive stoichiometric coefficients that are linked to the component X_j. The definitions of *reactive sum* and *product sum* vary according to the sign of the total concentration $[\tilde{T}_j]$ (negative total concentrations may arise with ion exchange, while null totals occur in the presence of H^+)

If $[\tilde{T}_j]\geq 0$

$$sum_j^{reac}=\sum_{a_{i,j}>0}a_{i,j}[C_i] \tag{33}$$

$$sum_j^{prod}=[\tilde{T}_j]-\sum_{a_{i,j}<0}a_{i,j}[C_i] \tag{34}$$

If $[\tilde{T}_j]<0$

$$sum_j^{reac}=|[\tilde{T}_j]|+\sum_{a_{i,j}>0}a_{i,j}[C_i] \tag{35}$$

$$sum_j^{prod}=-\sum_{a_{i,j}<0}a_{i,j}[C_i] \tag{36}$$

The PCF method is an empirical method. Once the equilibrium solution is found, the reactive sum equals the product sum. This method is another formulation for the conservation Eq. 2. Thanks to the repartition between the reactive materials and products, one can check if the component concentration value $[X_j^n]$ is too high (*product sum* greater than *reactive sum*) or too low (*reactive sum* lower than *product sum*). The component concentrations should then be updated according to formula (37), which increases or decreases the component concentration depending on the respective values of the reactive and product sums

$$[X_j]^{n+1}=[X_j]^n\left(\frac{sum_j^{prod,n}}{sum_j^{react,n}}\right)^{1/a_{i0,j}} \tag{37}$$

Nevertheless, formula (37) does not consider simultaneous changes in all the components' concentrations. A weight factor θ_j is introduced into Eq. 32 to avoid unfavorable oscillations. In this work, we exploit the same approach and structure by simply turning Eq. 32 into the following:

$$\xi_j^{n+1}=\left(1-\theta_j^n\right)\xi_j^n+\frac{\theta_j^n}{a_{i0,j}}\left(\ln\left(sum_j^{prod,n}\right)-\ln\left(sum_j^{reac,n}\right)\right) \tag{38}$$

Then, we define the weight factor θ_j as follows

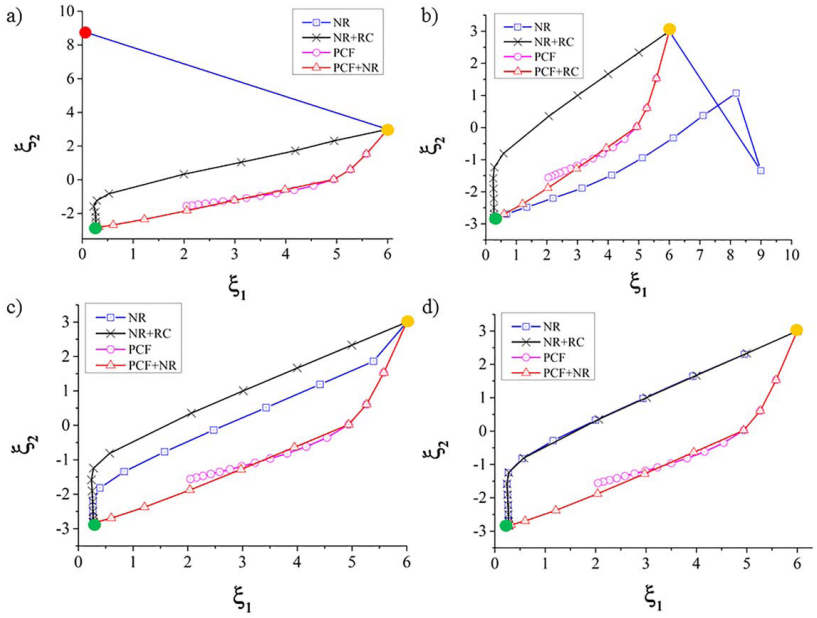


Figure 1. Effect of the round-off error 10^{-d} on the iterative solution.

(a–d) show the path to the solution of four different values of d, (d = 3, 4, 5, and 6 respectively). The yellow dots represent the initial condition, the green dots the solution and the red dot a unacceptable solution. [Color figure can be viewed at wileyonlinelibrary.com]

$$\begin{aligned} &if\left(sum_j^{reac} > sum_j^{prod}\right) then \quad \theta_j^n = \alpha - \frac{sum_j^{prod}}{sum_j^{reac}} \cdot \beta \\ &if\left(sum_j^{prod} > sum_j^{reac}\right) then \quad \theta_j^n = \alpha - \frac{sum_j^{reac}}{sum_j^{prod}} \cdot \beta \end{aligned} \tag{39}$$

where α and β are constants that control the amplitude of the step toward the solution. The step should be allowed to be large at the beginning (when sum_j^{reac} and sum_j^{prod} are different) but should be limited near the solution (when the two quantities are similar). In this work, these constants are set at α=0.1 and β=0.08. The weighting factor is updated only if $\theta_j^n > \theta_j^{n-1}$.

Numerical Experiments

A first simplified numerical example is proposed to highlight the effects of matrix conditioning and round-off errors on chemical equilibria solutions. The simplified chemical system is composed of three species and two components and is detailed in Table 1. The system is solved by the NR method, NR with Row-Column scaling (NR-RC), the PCF method, and PCF coupled with NR. For the initial conditions that are defined in Table 1, the associated matrix for the NR method has a condition number of 1.4×10^3 and the scaled matrix (NR-RC) has a condition number of 9.3×10^2. These condition numbers are defined in Eq. 17 by using the infinity norm. As shown by Eq. 16, the accuracy of the solution depends on the condition number and the round-off errors. Figure 1 shows the effects of the round-off errors on the path from the initial guesses to the solution of the above system when the NR method is used. Round-off errors are obtained by the truncation of numerical values to a selected number of digits.

The paths throughout the solution of four alternatives of the same simplified problem are shown in Figure 1 for different values of d, the exponent of the round-off error (d = 3, 4, 5, and 6 for Figures 1a–d, respectively). Since the problem has only two unknowns (components ξ_1 and ξ_2), the paths from the common initial guesses (ξ_1 =6.0 and ξ_2 =3.0) to the solution (ξ_1 = 0.3 and ξ_2 = −2.9) can be represented on a 2D graph.

In the worst case (round-of error 10^{-3}), the method does not converge (Figure 1a) or converges after some amount of iterations, even if the system is not accurately solved during the first iterations (Figures 1b, c). Once an acceptable accuracy is defined (round-of error 10^{-6}), NR and NR + RC lead to very similar solutions (Figure 1d). NR + RC is much less sensitive to the accuracy because the matrix is better conditioned, even if the contrast in the condition numbers is not very high for this example. The path to the solution does not change for the studied round-off errors. PCF is a first-order method that does not require a system to be solved. The method is provided here to illustrate its advantages and drawbacks; specifically, the method is robust far from the solution but inefficient in the neighborhood of the solution (the computation is stopped after 15 iterations for this example). Finally, the association of PCF

and NR provides an interesting alternative. The algorithm starts with a prescribed number of PCFs (three in this example) and switches to NR to reach the solution. This association is not sensitive to the round-off errors in this example. Finally, NR may reach the solution in fewer iterations than NR + RC even if the system matrix is not solved properly (Figure 1c). The wrong direction of descent for the first iteration is more efficient than the correct one. This process may occur a very few times and is not reliable.

More realistic problems involve a larger number of species and components than those involved in the previous experiment. The efficiencies of the different algorithms are studied with seven test cases of increasing complexity. Each test case is solved through the technique clarified in the previous sections: the concentrations of all the chemical species C_i in the system are computed based on a set of fixed components X_j. The majority of the test cases are available in the literature and considered to be numerically challenging because of the large ranges of stoichiometric coefficients and equilibrium constants. The number of species and components for each test case are summarized in Table 2.

The first and simplest test case is Gallic Acid, a system proposed by Brassard and Bodurtha[57] as an example of the onsets of problems in numerical methods. The system was originally studied in relation to Al(III) speciation in natural waters.[58] This first test case is characterized by the presence of 17 chemical species that can be described through the combination three components. Since no solid phase is taken into consideration here, the Jacobian is a 3×3 matrix, the smallest of the whole set of test cases. Also the range of variation of equilibrium constants is the smallest of those examined. Increasing complexity is found in MoMaS Easy test case, a synthetic benchmark designed to evaluate the performances of computational codes and published in a special issue of Computational Geosciences.[59] This test case is characterized by the presence of 12 species and the number of base components required to describe the system is 5. The range of equilibrium constants is higher (47 against 35 orders of magnitude) than in the previous case. Complexity increases significantly approaching Pyrite[31] test cases. This example describes the environment for the potential precipitation of Pyrite (FeS_2). In a first variant of the test case, precipitation of a solid phase is denied and the system is composed of 40 dissolved species described through four components. The number of components is limited but the differences between equilibrium constants are huge (see Table 2). A second variant of the test case, Pyrite Mineral, is a copy of the previous enriched with the possible formation of three minerals ($Fe, FeSO4, FeS_2$). The size of the Jacobian matrix becomes 7×7. Continuing in the presentation of test cases, it becomes harder to precisely assess the order of complexity. MoMaS Hard test case comes from the same set of synthetic examples of MoMaS Easy.[59] It's characterized by the presence of 17 chemical species and described through six components. Two minerals are tested for precipitation, making the Jacobian a 8×8 matrix while the range of equilibrium constants remains the same of MoMaS Easy. Two test cases involving iron and chrome are also studied. Fe Cr test case is the simplest and describes a chemical system of 40 species and seven components without precipitate. In this case, the number of components is higher than in the Pyrite test cases but the range of variation of equilibrium constants is considerably smaller. Fe Cr Min is a variant of the previous case in which three minerals are tested for precipitation, making the Jacobian matrix size 10×10.

Table 1. Stoichiometric Coefficients, Thermodynamic Constant (K), Totals of the Components ξ_1 and ξ_2 (T), and Logarithm of the Initial Activity ($\xi_{Initial}$)

	ξ_1	ξ_2	K
ξ_1	1	0	
ξ_2	1	1	
ζ_1 ξ_2	−1	3	2000.0
T	1.0	1.0	
$\xi_{Initial}$	6.0	3.0	

The whole sets of equilibrium constants and a complete description of the chemical species and components are presented in Appendix B.

Because the motivation of this work is to evaluate the behavior of the NR method and associated algorithms, we only compute the first solution of the chemical system, i.e., if precipitation occurs and no precipitation was assumed, the computation is not repeated. Therefore, we also assumed that the activity coefficients were constant and equal to one. This approximation reduces the non-linearity of the problem but will not affect our conclusions.

Usually, initial guesses for the NR method are chosen with great care (i.e., from the concentrations at the previous time step in transient computations). However, initial guesses are not always known, especially for the first time step or for sharp fronts, where the concentrations show abrupt changes from cell to cell in the reactive transport code. Because the aim of this work is to test the robustness of the method no matter the initial conditions, these conditions are chosen randomly in a prescribed interval following a uniform distribution. The boundaries of the intervals are estimated as described and explained in Appendix C. The robustness of the procedure is also enhanced by prescribing boundaries to the NR increments (Appendix C). We searched the solutions of chemical equilibria for a single problem (combination test case/method of solution) using 30,000 different initial component concentrations X_j, except for Gallic Acid (10,000), which appeared to be the easiest test case. We assumed that 30,000 (10,000 for Gallic Acid) simulations adequately represent the behaviors of the convergence rates for each example because the percentage of failure (actually the mean number of iterations and the standard deviation of the number of iterations that is required to converge) remains unchanged after 25,000 runs at most.

The different resolution methods are: regular NR method (called No Scaling), NR method implemented with the four scaling techniques (RC, RI, sDsD, DI) and with the MEq procedure and regular NR coupled with PCFs implemented when the condition numbers is greater than a user's defined value (PCF).

The convergence criterion is written based on the residual $Y_j = [\tilde{T}_j] - [T_j]$ and the quantity $W_j = [\tilde{T}_j] + \sum_i |a_{i,j}[C_i]|$. The totals that appear in the previous equations were defined in Eq. 9. In the presence of precipitates, W_j is not computed and the value of the convergence criterion corresponds to the residual

$$\begin{aligned} err_j &= \frac{Y_j}{W_j} \quad j=1,\ldots,N_X \\ err_j &= Y_j \quad j=N_X+1,\ldots,Nx+Ncp \end{aligned} \tag{40}$$

The NR iterations end when the convergence criterion is satisfied, i.e., when the highest residual is lower than a given

Table 2. Summary of the Test Cases

	Dissolved species	Adsorbed species	Precipitates	Components	$\log_{10}K_{min}$–$\log_{10}K_{max}$
Gallic acid	17	–	–	3	−39.56 to −4.15
MoMaS easy	9	3	–	5	−12 to 35
Pyrite	40	–	–	4	−520.6 to 19.17
Fe Cr	40	–	–	7	−83.17 to 80.9
Pyrite M	40	–	3	4	−520.6 to 19.17
MoMaS hard	12	3	2	6	−12 to 35
Fe Cr Min	40	–	3	7	−83.17 to 80.9

threshold ($tol=10^{-12}$). The Newton iterations are also stopped when the maximum number of iterations exceeds 2000.

When possible (with symmetric matrices), the evaluation of the condition number is performed with two approaches: (1) the absolute value of the ratio between the highest and lowest eigenvalues of the Jacobian matrix (18) and (2) the product of the norm one of matrix $\mathbf{J}$ and its inverse $\mathbf{J}^{-1}$ (17). The subroutine DE3LRG in the IMSL library[60] is used to evaluate the eigenvalues.

The linear systems (13) and/or (27) within the NR procedure are solved through LU factorization with quadruple precision.

The MEq algorithm performs scaling on the Jacobian matrix until the infinity norm of each RC equals one. We set a maximum number of iterations $n^{R}_{max}=5$ to avoid excessive slowdowns in the computation.

The algorithm starts with PCFs when implemented and depending on the condition number, i.e., the algorithm is activated if the condition number is higher than a fixed threshold. Only 10 iterations of this method are performed because the aim is to use the algorithm as a type of preconditioner and not to reach the solution (also in light of the results in Figure 1).

Numerical Simulations: Discussion

The purpose of scaling is to reduce the condition number of a linear system. Scaling procedures consist of changing the coefficient in the matrix and the right hand side of the system of equations accordingly. We evaluated the condition number of the Jacobian matrix of the nonlinear system before and after scaling. We presented the results for problems with condition numbers at different orders of magnitude and for a round-off error of 10^{-32}.

When the condition number was on the order of 10^{60} or lower (as in Figure 2a), both methods of computing the condition number (as the product of the norm one of the Jacobian matrix and its inverse [Eq. 17] and as the ratio of the eigenvalues [Eq. 18]) were effective and provided the same results. Under this circumstance, the condition number was generally reduced by RC scaling, even if the amplitude of this reduction varied widely.

When the Jacobian matrix had a condition number on the order of 10^{200} or more (Figure 2b), the estimation of the condition number was no longer reliable: the inversion of the Jacobian matrix was heavily imprecise, and the DE3LRG subroutine from IMSL failed to estimate the eigenvalues for condition numbers that were greater than 10^{50}. RC scaling was effective except for condition numbers that were approximately $10^{120}-10^{150}$, which shows that scaling may be useful but not universally so. Figure 2 shows that the condition number decreased with the distance from the solution. When a large number of iterations were required to converge, the initial values of the condition number were extreme. When a modest number of iterations were necessary to reach the solution, the condition number had smaller values and showed more regular behavior.[46]

When the initial guesses are far from the solution, the solutions obtained during the first steps of the iterative procedure may not be accurate which will lead to a poor estimate of the direction of descent. However, because the Jacobian matrix is updated after each iteration, it will probably lead to additional iterations but will not affect the accuracy of the solution obtained after convergence.

The results are presented through the relative number of obtained solutions as a function of the number of iterations for each test case (Figures 3–9). When the solution is obtained for all the initial conditions ($N_{tot}=30{,}000$ for all test cases except Gallic Acid), the relative number equals 1. The distributions of the initial condition numbers are also provided in Figures 3–9. Table 3 provides the number of iterations that are required to solve 50, 70, and 90% of the problems for each test

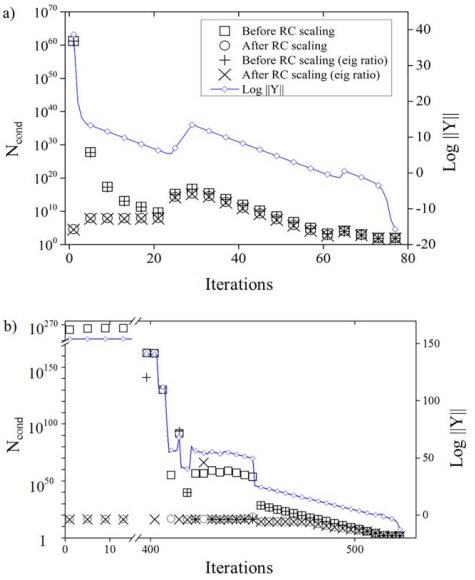


Figure 2. Variation in condition numbers throughout the minimization for the Pyrite test case with favorable initial guesses (a) and poor initial guesses (b) of the component concentrations.

[Color figure can be viewed at wileyonlinelibrary.com]

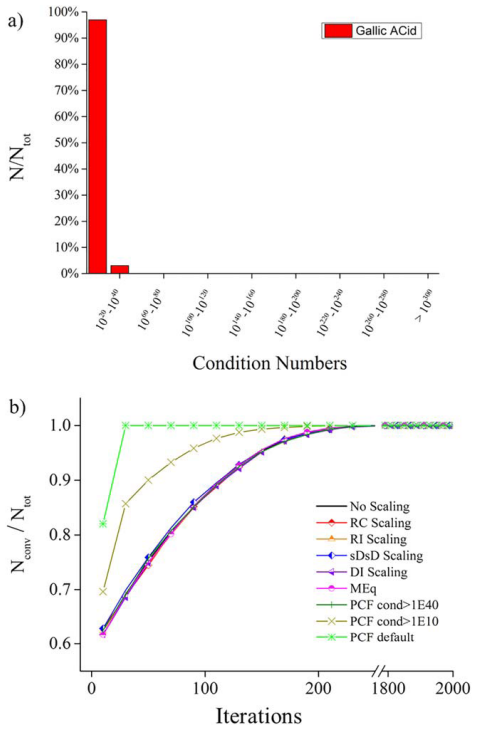


Figure 3. Initial condition number and relative number of solutions for the Gallic Acid test case.

[Color figure can be viewed at wileyonlinelibrary.com]

case and each algorithm. The number of failures (solution not reached after 2000 iterations) is listed in Table 4.

The results vary widely between test cases, particularly for the implementation of scaling techniques.

In the case of **Gallic Acid** speciation (Figure 3), no differences existed among the curves that represented the standard NR method, NR with scaling techniques or PCF, which were activated when the condition number was greater than 10^{20}, because most of the condition numbers were below 10^{20} (Figure 3a). Because the condition numbers were smaller than the threshold (10^{32}) that complicates the computation of reliable digits in terms of the round-off error (10^{-32}), the system was solved accurately with or without scaling. When PCFs were performed by default, the number of iterations that were required to solve 100% of the problems dropped from 250 to approximately 30. The results that correspond to the activation of PCFs when the condition number was greater than 10^{10} provide an intermediate result because some computations were run without PCFs (condition number less than 10^{10}).

The distribution of the initial condition numbers reached 10^{90} for the **MoMaS Easy** test case (Figure 4a). NR when implemented with scaling performed better than standard NR except for DI (Figure 4b). 90% of the problems were solved within 140 iterations with the matrix equilibrium technique where other scaling techniques required more than 200 (Table 3). The solution of NR coupled with PCFs constantly outperformed the integration of scaling (Tables 3 and 4). This test case also demonstrated the main shortcoming of the PCF method: as shown in Figure 4, PCFs can slow down the solution process in the neighborhood of the solution. When PCFs were applied by default (i.e., regardless if they were needed), almost no solutions were obtained within 20 iterations and only a few within 40. This result confirms that the NR method needs no strengthening when the initial conditions are favorable.

The difficulty in finding a solution increased for the **Pyrite** test case, which exhibited numerous condition numbers that were greater than 10^{200} (Figure 5a). Approximately 85% of the problems were solved by NR or NR with scaling with the same efficiency, except the MEq technique which was able to solve 90% of the problems in about 1800 iterations (Table 3) and solved 95% of the problems (Table 4). The increase in the number of iterations (from approximately 200–600) did not change the probability of success in solving the problem. This probability slightly increased after 1000 iterations, but none of the NR methods could solve all the problems. Scaling by MEq was the most efficient scaling technique for this example. The PCFs appeared to be very efficient, solving all the problems with about 50 iterations.

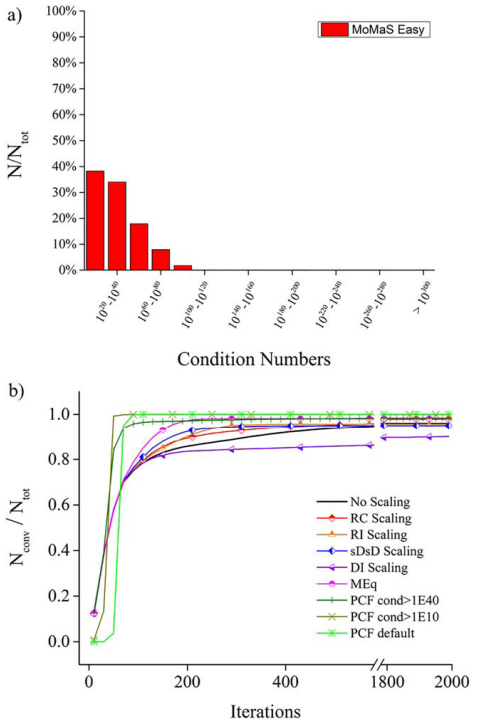


Figure 4. Initial condition number and relative number of solutions for the MoMaS Easy test case.

[Color figure can be viewed at wileyonlinelibrary.com]

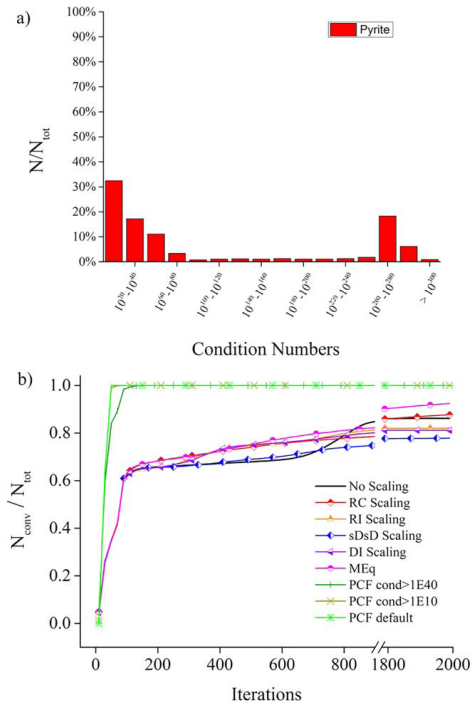


Figure 5. Initial condition number and relative number of solutions for the Pyrite test case.

[Color figure can be viewed at wileyonlinelibrary.com]

Most of the condition numbers of the standard NR ranged from 10^{40} to 10^{160} for the **Fe Cr** test case (Figure 6a). Scaling did not improve the convergence probability (Figure 6b and Table 3) for that range of condition numbers. The scaling techniques improved the system's matrix properties but not enough to compute an accurate solution of the system. Therefore, scaling was inefficient in this case. Moreover, no significant differences existed in the activation of PCFs after different thresholds of condition numbers, which means that the large majority of the problems had a condition number that was greater than 10^{40}.

Pyrite Mineral, exhibited high contrasts in the condition numbers, with 1/3 of the system matrices having a condition number that was greater than 10^{300} (Figure 7a). At this level of complexity, scaling techniques were not efficient enough to reduce the high condition numbers. Condition numbers that were greater than 10^{300} were indeed reduced but still remained enormous (10^{260}–10^{280}) (Figure 7a). Therefore, NR with scaling was as efficient as NR without scaling at solving this system of equations (Figure 7b). Again, the PCFs appeared to be very efficient, and 90% of the problems were solved in fewer than 50 iterations (Table 3).

The **MoMaS Hard** test case contemplated the presence of precipitates, and its condition numbers ranged from 10^{10} to 10^{200} (Figure 8a). Therefore, some problems were solved with a small number of iterations (less than 50 for more than 50% of the initial conditions—Table 3), except for the MEq. For more difficult problems, the implementation of scaling techniques (except MEq) improved the situation slightly. MEq reduced the condition numbers significantly (Figure 8a) and the number of failures more than other scaling techniques but made the process of convergence slower (Figure 8b). The implementation of PCFs coupled with the standard NR method improved in a more noticeable way both the robustness and speed of the convergence. Again, the standard PCFs required more iterations than the adapted PCFs to solve the easiest problems.

For the **Fe Cr Min** test case, the addition of precipitates considerably changed the matrix properties compared to **Fe Cr** (Figure 9a). All scaling techniques dealing with both rows and columns reduced the number of failures and improved the robustness of the algorithm (Figure 9b). For this example, the sDsD scaling method was the most appropriate. At this level of complexity, some scaling techniques (RI and DI scaling) were inefficient and, in particular, less efficient than NR without scaling at solving this system of equations. Even if the PCFs provided a further improvement in terms of speed and robustness, nearly the half of problems remained unsolved (Table 4).

The effects on CPU time were negligible for all the scaling procedures with the exception of MEq, which increased the duration of a single iteration by approximately 20%. However, this significant increase in the CPU time for one iteration is

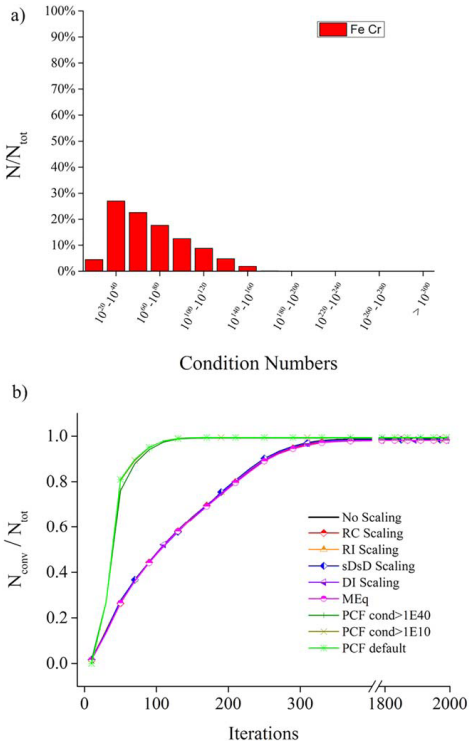


Figure 6. Initial condition number and relative number of solutions for the Fe Cr test case.

[Color figure can be viewed at wileyonlinelibrary.com]

compensated by the total number of iterations that are required to solve a given percentage of the problems (Table 3). Solving 70% of the problems for the Pyrite test case with the standard NR method required approximately 700 iterations. Conversely, only 350 iterations were necessary for the NR method with MEq to solve the same percentage of problems. This same situation occurred for the MoMaS Easy test case, where solving 90% of the problems with standard NR required approximately double the iterations of NR with MEq. The computational effort for PCFs for one iteration was significantly smaller than that for any NR method because this method did not require the computation of the Jacobian matrix and its solution.

The effects of coupling the NR method with scaling techniques or PCFs on the robustness were evaluated by counting the number of failures (i.e., non-convergence events within 2000 NR iterations) while searching for the solution. The percentages of failures (computed with respect to 30,000 attempts for the test cases) are reported in Table 4. While PCFs improved the robustness of the code without exception, the outcomes of implementing scaling techniques strongly depended on the test case. The MoMaS Easy test reduced the number of failures when RC scaling and MEq were applied, while other techniques were counterproductive in terms of robustness (if the limit of the NR number of iterations was set to 2000). In the Pyrite test case, RC scaling and MEq

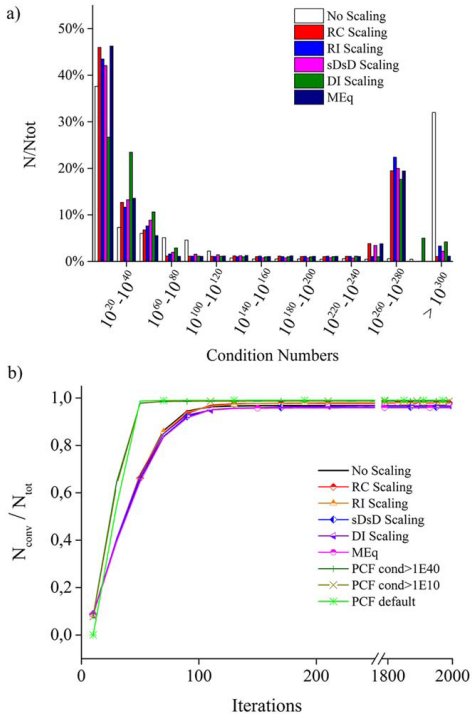


Figure 7. Initial condition number and relative number of solutions for the Pyrite Mineral test case.

[Color figure can be viewed at wileyonlinelibrary.com]

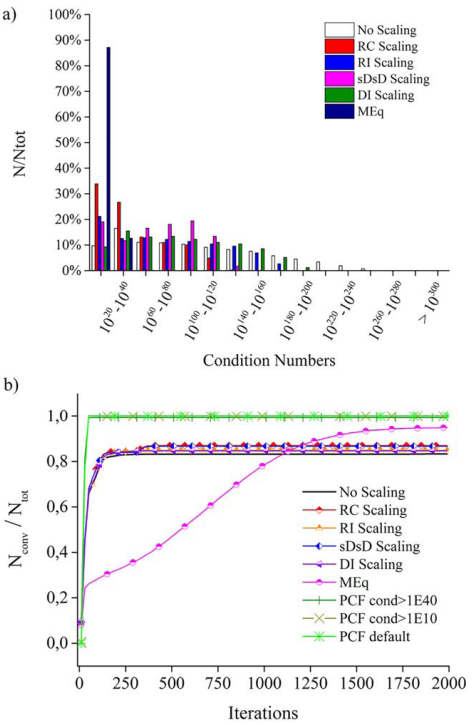


Figure 8. Initial condition number and relative number of solutions for the MoMaS Hard test case.

[Color figure can be viewed at wileyonlinelibrary.com]

increased the robustness of the method, while other scaling procedures induced the opposite effect.

For the MoMaS Hard test case, very small increments of robustness were registered for RI scaling, MEq and DI scaling. All the preconditioners (but DI scaling) for the Fe Cr Min test case seemed to significantly reduce the number of failures. A global look at the bottom half of Table 4 suggests that PCFs drastically outperformed scaling techniques in terms of reducing the number of failures after 2000 NR iterations. Among the scaling techniques, MEq and RC scaling worked the best, significantly reducing the number of non-convergences in some cases.

Conclusions

The results of scaling strongly depended on the test case. In problems such as the Pyrite, MoMaS Easy or MoMaS Hard test cases, scaling techniques (especially Row-Column scaling and MEq) generally exerted a positive impact. Meanwhile, the utility of scaling was questionable for other problems, such as the Fe Cr Min and Pyrite Mineral test cases. These results suggest three possible classifications:

- "Easy" problems, such as the Gallic Acid test case, which are insensitive to scaling because NR without scaling is efficient. The reduction in the condition numbers by scaling does not improve the efficiency of NR because the

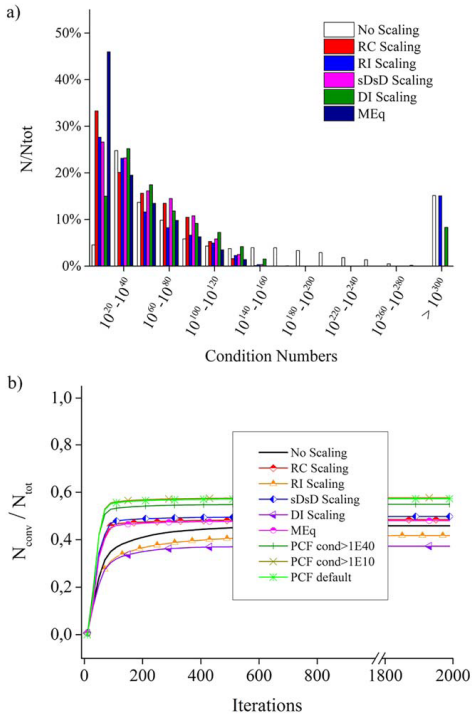


Figure 9. Initial condition number and relative number of solutions for the Fe Cr Mineral test case.

[Color figure can be viewed at wileyonlinelibrary.com]

Table 3. Number of Iterations that are Required to solve 50, 70, and 90% of the Studied Problems for each Test Case (10, 30, and 50% for Fe Cr Min)

	(%)	No scaling	RC scaling	RI scaling	MEq	PCF def
Gallic acid	50	20	20	20	20	<20
	70	40	40	40	40	<20
	90	120	120	120	120	<20
MoMaS E.	50	50	50	50	50	75
	70	80	80	80	80	<80
	90	340	240	220	140	<80
Pyrite	50	90	90	90	90	<40
	70	680	300	340	320	40
	90	>2000	>2000	>2000	1800	<50
Fe Cr	50	120	120	120	120	50
	70	180	180	180	180	60
	90	260	260	260	260	80
Pyrite M.	50	45	50	50	50	40
	70	60	60	60	65	45
	90	90	90	90	95	50
MoMaS H.	50	45	45	45	560	35
	70	85	70	80	860	40
	90	>2000	>2000	>2000	1330	45
Fe Cr Min	10	35	35	35	40	35
	30	75	55	90	55	45
	50	>2000	>2000	>2000	>2000	75

Table 4. Failure % of the Different Algorithms for each Test Case (NB: no failure for Gallic Acid and FeCr Test Cases)

	MoMaS E	Pyrite	Pyrite M	MoMaS H	Fe Cr Min
No scaling	4.1	13.8	3.3	16.8	54.2
RC	2.1	12.3	2.2	13.1	51.5
RI	4.4	17.9	2.0	15.1	58.3
sDsD	5.0	22.1	3.8	13.3	50.3
DI	10.0	18.8	3.2	15.4	62.8
MEq	1.9	7.5	4.2	5.2	52.0
PCF cond $>10^{40}$	1.7	0.0	1.4	0.7	45.1
PCF cond $>10^{10}$	0.1	0.0	1.4	0.0	43.0
PCF def	0.0	0.0	1.0	0.0	43.0

initial condition numbers are already small enough (i.e., smaller than the round-off error);

- Problems where scaling reduces the condition numbers to values on the order of magnitude of the round-off error;
- Problems that are very difficult to solve and where scaling does not reduce the condition numbers sufficiently, providing unpredictable reactions to scaling. This was the case when precipitation were included (three last test cases).

Among the scaling techniques, MEq and RC appeared to be globally useful in terms of increasing the robustness of the algorithm and sometimes reducing the number of iterations that were required to reach the solution. However, the drawback of this scaling method was the additional computational costs.

The coupling of NR with PCFs produced results that were quite similar for every test case. This coupling drastically reduced the number of failures, the number of iterations that were required to reach convergence, and their variability (the cumulated frequency curves are nearly vertical). However, PCFs alone should be avoided because they are very slow when they approach the solution.

The previous results seem to indicate that the ill-conditioned linear systems that arise in NR iterations are only an obstacle to fast convergence for problems that are not too easy or tough. When the condition numbers are too high, they are no longer the cause of an eventual failure and become a symptom of the NR method's inadequacy. PCFs bring the values of the unknowns close to the solution, where the NR algorithm is known to be extremely efficient. This coupling seems to bypass the problem of ill-conditioned linear systems, transforming tough problems into more easy problems and avoiding the intermediate zone. This behavior enforces the idea that the condition numbers of the Jacobian matrix decrease while approaching the solution.

At this stage of our work, we strongly recommend coupling PCFs with NR. PCFs should only be activated when needed, i.e., when the condition number of the Jacobian matrix is greater than a given threshold. Because the computation of the condition number might require some computational costs, this process can be replaced by the norm of the residuals and a corresponding user's defined threshold. Moreover, the Jacobian matrix can be scaled by the RC method, which is the first iteration of the MEq scaling technique, to improve the robustness of NR without increasing the CPU time.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. The first author acknowledges support by the "Initiative d'Excellence" from Strasbourg University, France.

Notation

C_i = = chemical species i, –
X_j = = chemical component j, –
Cp_l = = precipitate l, –
N_C = = number of chemical species, –
N_x = = number of chemical components, –
N_{Cp} = = number of precipitates, –
N_t = = number of unknowns $N_t = N_x + N_{Cp}$, –
$[\bullet]$ = = concentration, mol/L
$\{\bullet\}$ = = activity, –
ξ_j = = $\xi_j = \ln(\{X_j\})$ generic unknown of the system in logarithmic form, –
$b_{i,j}$ = = stoichiometric coefficient in the composition of species C_i for component X_j in conservation law, –
$bp_{l,j}$ = = stoichiometric coefficient in the composition of mineral Cp_l for component X_j in mass action law, –
K_i = = constant of thermodynamic equilibrium for the formation of species C_i, –
K_s^l = = solubility product of mineral l, –
$[T_j]$ = = total concentration of component X_j, mol/L
$[\overline{T_j}]$ = = total prescribed concentration of component X_j, mol/L
γ_i = = activity coefficient related to chemical species i, mol/L
Y_j = = residue between measured and computed conserved total concentrations, mol/L
$\kappa(\mathbf{J})$ = = condition number of the Jacobian matrix $\mathbf{J}$, –
$\mathbf{D_1}, \mathbf{D_2}$ = = generic diagonal matrix
$\mathbf{a}_i$ = = generic row in the Jacobian matrix $\mathbf{J}$
$\mathbf{c}_i$ = = generic column in the Jacobian matrix $\mathbf{J}$

Abbreviations

NR = = Newton Raphson
PCF = = positive continuous fractions
RC = = row column scaling
RI = = row identity scaling
DI = = diagonal identity scaling
sDsD = = square diagonal scaling
MEq = = matrix equilibration

Literature Cited

1. Saaltink MW, Carrera J, Ayora CA. Comparison of two approaches for reactive transport modeling. *J Geochem Explor*. 2000;69:97–101.
2. Saaltink MW, Carrera J, Ayora C. On the behavior of approaches to simulate reactive transport. *J Contam Hydrol*. 2001;48(3):213–235.
3. Steefel CI, Appelo CAJ, Arora B, Jacques D, Kalbacher T, Kolditz O, Lagneau V, Lichtner PC, Mayer KU, Meeussen JCL, Molins S, Moulton D, Shao H, Šimůnek J, Spycher N, Yabusaki SB, Yeh GT. Reactive transport codes for subsurface environmental simulation. *Comput Geosci*. 2015;19(3):445–478.
4. Lichtner PC, Hammond GE, Lu C, Karra S, Bisht G, Andre B, Mills R, Kumar J. PFLOTRAN user manual. A massively parallel reactive flow and transport model for describing surface and subsurface processes. 2015. Available at: https://bytebucket.org/pflotran/pflotran-dev/wiki/Documentation/Files/user_manual.pdf?rev=2d64a0339b289e2e44387609f0f32bd78c77a04f
5. Mayer KU, Frind EO, Blowes DW. Multicomponent reactive transport modeling in variably saturated porous media using a generalized formulation for kinetically controlled reactions: reactive transport modeling in variably saturated media. *Water Resour Res*. 2002;38(9):13-1–13-21.
6. Henderson TH, Mayer KU, Parker BL, Al TA. Three-dimensional density-dependent flow and multicomponent reactive transport modeling of chlorinated solvent oxidation by potassium permanganate. *J Contam Hydrol*. 2009;106(3–4):195–211.
7. Jacques D, Šimůnek J. User manual of the multicomponent variably-saturated flow and transport model HP1. Description verification and examples. Version 1.0, SCK-CEN-BLG-998, Waste and Disposal, SCK-CEN, Mol, Belgium, 2005:1–79.
8. Šimůnek J, Jacques D, Šejna M, van Genuchten M. The HP2 program for HYDRUS (2D/3D): a coupled code for simulating two-dimensional variably-saturated water flow, heat transport, and biogeochemistry in porous media, Version 1.0, PC Progress, Prague, Czech Republic, 2012:1–76.
9. Simunek J, Jacques D, Langergraber G, Bradford SA, Šejna M, van Genuchten MT. Numerical modeling of contaminant transport using HYDRUS and its specialized modules. *J Indian Inst Sci*. 2013;93(2):265–284.
10. Prommer H, Post V. PHT3D, a reactive multicomponent transport model for saturated porous media. User's manual v2.10, Available at: http://www.pht3d.org/pht3d_exe.html.
11. van der Lee J, De Windt L, Lagneau V, Goblet P. Module-oriented modeling of reactive transport with HYTEC. *Comput Geosci*. 2003;29(3):265–275.
12. Xu T, Pruess K. Modeling multiphase non-isothermal fluid flow and reactive geochemical transport in variably saturated fractured rocks: 1. Methodology. *Am J Sci*. 2001;301(1):16–33.
13. Xu T, Sonnenthal E, Spycher N, Pruess K. TOUGHREACT. A simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media: applications to geothermal injectivity and CO2 geological sequestration. *Comput Geosci*. 2006;32(2):145–165.
14. Xu T, Spycher N, Sonnenthal E, Zhang G, Zheng L, Pruess K. TOUGHREACT Version 2.0: a simulator for subsurface reactive transport under non-isothermal multiphase flow conditions. *Comput Geosci*. 2011;37(6):763–774.
15. Xu T, Sonnenthal E, Spycher N, Zheng L. TOUGHREACT V3.0-OMP reference manual: a parallel simulation program for non-isothermal multiphase geochemical reactive transport. Report LBNL-draft, Lawrence Berkeley National Laboratory, Berkeley, CA, 2014.
16. White MD, Oostrom M. STOMP subsurface transport over multiple phases version 4.0 user's guide. PNNL-15782, Pacific Northwest National Laboratory, Richland, WA, 2006.
17. White MD, McGrail BP. Stomp subsurface transport over multiple phase version 1.0 addendum: eckechem equilibrium-conservation-kinetic equation chemistry and reactive transport. PNNL-15482, Pacific Northwest National Laboratory, Richland, WA, 2005.
18. Steefel CI. CrunchFlow—software for modeling multicomponent reactive flow and transport, User's Manual, 2009. Available at: http://www.csteefel.com/CrunchFlowManual.pdf.
19. Yeh GT, Tsai CH. HYDROGEOCHEM 6.0 a two-dimensional model of coupled fluid flow, thermal transport, HYDROGEOCHEMical transport, and geomechanics through multiple phase systems Version 6.0 (FACTM2D a model for multi-phase flow analysis and reactive chemical transport, thermal transport, and mechanics simulation, 2-dimensional version)—theoretical basis and numerical approximation. Graduate Institute of Applied Geology, National Central University, Jhongli, 2013.
20. Tsai CH, Yeh GT, Ni CF. HYDROGEOCHEM 6.0: a model to couple thermal-hydrology-mechanics-chemical (THMC) processes USER GUIDE. Graduate Institute of Applied Geology, National Central University, Jhongli, 2013.
21. Parkhurst DL, Appelo CAJ. User's guide to PHREEQC (Version 2): a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations Water-Resources Investigations, Report 99–4259, Denver, CO, USA, 1999.
22. van der Lee J. Thermodynamic and Mathematical concepts of CHESS. Report LHM/RD/98/39, Ecole Nationale Superieure des Mines de Paris, Paris, 1998.
23. van der Lee J, De Windt L. Chess tutorial and cookbook, updated for version 3.0. Report LHM/RD/02/13, Ecole Nationale Superieure des Mines de Paris, Paris, 2002.
24. Kolditz O, Bauer S, Bilke L, Böttcher N, Delfs JO, Fischer T, Gorke UJ, Kalbacher T, Kosakowski G, McDermott CI, Park CH, Radu F, Rink K, Shao H, Shao HB, Sun F, Sun YY, Singh AK, Taron J, Walther M, Wang W, Watanabe N, Wu Y, Xie M, Xu W, Zehner B. OpenGeoSys: an open-source initiative for numerical simulation of thermo-hydro-mechanical/chemical (THM/C) processes in porous media. *Environ Earth Sci*. 2012;67(2):589–599.
25. Bea SA, Carrera J, Ayora C, Batlle F, Saaltink MW. CHEPROO: a Fortran 90 object-oriented module to solve chemical processes in Earth Science models. *Comput Geosci*. 2009;35(6):1098–1112.
26. Allison JD, Brown DS, Kevin J. MINTEQA2/PRODEFA2: A geochemical assessment model for environmental systems: version 3.0 user's manual. Environmental Research Laboratory, Office of Research and Development, US Environmental Protection Agency Athens, GA, 1991.
27. Wood JR. Calculation of fluid-mineral equilibria using the simplex algorithm. *Comput Geosci*. 1993;19(1):23–39.
28. Buzzi-Ferraris G, Manenti F. Nonlinear systems and optimization for the chemical engineer. Solving numerical problems. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2014.

29. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd ed. Vol. 1. Cambridge: Cambridge University Press, 1997.
30. Chen K. *Matrix Preconditioning Techniques and Applications*. Cambridge: Cambridge University Press, 2005.
31. Hoffmann J. Reactive transport and mineral dissolution/precipitation in porous media: efficient solution algorithms, benchmark computations and existence of global solutions. PhD Thesis, Universität Erlangen–Nürnberg, 2010.
32. Golub GH, Van Loan CF. *Matrix Computations*, 3rd ed. Baltimore: John Hopkins University Press, 1996.
33. Carrayrou J, Mosé R, Behra P. New efficient algorithm for solving thermodynamic chemistry. *AIChE J*. 2002;48(4):894–904.
34. Kulik DA, Wagner T, Dmytrieva SV, Kosakowski G, Hingerl FF, Chudnenko KV, Berner UR. GEM-Selektor geochemical modeling package: revised algorithm and GEMS3K numerical kernel for coupled simulation codes. *Comput Geosci*. 2013;17:1–24.
35. Leal AMM, Blunt MJ, LaForce TC. Efficient chemical equilibrium calculations for geochemical speciation and reactive transport modelling. *Geochim Cosmochim Acta*. 2014;131:301–322.
36. Reed MH. Calculation of multicomponent chemical equilibria and reaction processes in systems involving minerals, gases and an aqueous phase. *Geochim Cosmochim Acta*. 1982;46:513–528.
37. Westall C, Zachary LJ, Morel FMM. *Mineql: A Computer Program for the Calculation of Chemical Equilibrium Composition of Aqueous Systems*. Cambridge: Ralph M. Parsons Laboratory, Department of Civil Engineering, Massachusetts Institute of Technology, 1976.
38. Cederberg GA, Street RL, Leckie JO. A groundwater mass transport and equilibrium chemistry model for multicomponent systems. *Water Resour Res*. 1985;21:
39. Yeh GT, Tripathi VS. A model for simulating transport of reactive multispecies components: model development and demonstration. *Water Resour Res*. 1991;27(12):3075–3094.
40. Steefel CI, Lasaga AC. A coupled model for transport of multiple chemical species and kinetic precipitation/dissolution reactions with application to reactive flow in single phase hydrothermal systems. *Am J Sci*. 1994;294:529–592.
41. Steefel CI, MacQuarrie KTB. Approaches to modeling of reactive transport in porous media. *Rev Miner*. 1996;34:83–125.
42. Chilakapati A. RAFT: A simulator for reactive flow and transport of groundwater contaminants, PNL Rep. 10636, Pac. Northwest Lab., Richland, Wash. 1995.
43. Chilakapati A, Ginn T, Szecsody J. An analysis of complex reaction networks in groundwater modeling. *Water Resour Res*. 1998;34(7): 1767–1780.
44. Fang Y, Yeh G-T, Burgos WD. A general paradigm to model reaction-based biogeochemical processes in batch systems: a general paradigm to model biogeochemical processes. *Water Resour Res*. 2003;39(4):1083
45. Leal AMM, Blunt MJ, LaForce TC. A chemical kinetics algorithm for geochemical modelling. *Appl Geochem*. 2015;55:46–61.
46. Machat H, Carrayrou J. Comparison of linear solvers for equilibrium geochemistry computation. *Accepted Comput Geosci*. 2016.
47. Kiusalaas J. *Numerical Methods in Engineering with MATLAB*. New York: Cambridge University Press, 2005.
48. Chapra SC, Canale RP. *Numerical Methods for Engineers*, 7th ed. New York: McGraw-Hill Education, 2015.
49. Buzzi-Ferraris G. New trends in building numerical programs. *Comput Chem Eng*. 2011;35(7):1215–1225.
50. Wolery TW, Jarek RL. Software user's manual. EQ3/6, Version 8.0. Sandia National Laboratories, Albuquerque, 2003.
51. Marquardt DW. An algorithm for least square estimation of nonlinear parameters. *SIAM J Ind Appl Math*. 1963;11(2):431–441.
52. Knight PA, Ruiz D, Uçar B. A symmetry preserving algorithm for matrix scaling. *SIAM J Matrix Anal Appl*. 2014;35(3):931–955.
53. Bradley AM. Algorithms for the equilibration of matrices and their application to limited-memory Quasi-Newton methods. PhD Thesis, Stanford University, 2010.
54. Golub GH, Ortega JM. *Scientific Computing: An Introduction with Parallel Computing*. Boston, MA: Academic Press, 1993.
55. Delay F, Kaczmaryk A, Ackerer P. Inversion of interference hydraulic pumping tests in both homogeneous and fractal dual media. *Adv Water Resour*. 2007;30(3):314–334.
56. Ruiz D, Uçar BA. Symmetry preserving algorithm for matrix scaling. Report 7552 Institut National de Recherche En Informatique et en Automatique, Grenoble, 2011.
57. Brassard P, Bodurtha P. A feasible set for chemical speciation problems. *Comput Geosci*. 2000;26(3):277–291.
58. Öhman LO. Equilibrium studies of ternary aluminium (III) hydroxo complexes with ligands related to conditions in natural waters. Report University of Umeå, 1983.
59. Carrayrou J, Kern M, Knabner P. Reactive transport benchmark of MoMaS. *Comput Geosci*. 2009;14(3):385–392.
60. IMSL: Fortran Numerical Library. User's Guide, Math Library. Version 7.0. Viaual Numerics, Rogue Wave Software Company. Available at: http://www.roguewave.com/help-support/documentation/imsl-numerical-libraries.

Appendix A

The method to compute the Jacobian matrix is not affected by changing variables. We must compute the derivative of the system (9), which is written as a function of ξ with respect to ξ:

$$\frac{\partial Y_k}{\partial \xi_j} = \frac{\partial\left(\left[\tilde{T}_k\right] - [T_k]\right)}{\partial \xi_j} = -\frac{\partial T_k}{\partial \xi_j}. \tag{A1}$$

The known total concentrations $\left[\tilde{T}_k\right]$ are constant values. The derivative takes the following form:

$$\frac{\partial T_k}{\partial \xi_j} = \frac{\partial}{\partial \xi_j}\sum_{i=1}^{Nc} b_{i,k} \exp\left(\ln K_i + \sum_{j=1}^{Nx} b_{i,j}\xi_j\right). \tag{A2}$$

The derivative of a sum is the sum of the derivatives:

$$\frac{\partial T_k}{\partial \xi_j} = \sum_{i=1}^{Nc} b_{i,k}\left(\frac{\partial}{\partial \xi_j}\exp\left(\ln K_i + \sum_{j=1}^{Nx} b_{i,j}\xi_j\right)\right). \tag{A3}$$

The derivative of an exponential function is again an exponential function:

$$\frac{\partial T_k}{\partial \xi_j} = \sum_{i=1}^{Nc} b_{i,k}\exp\left(\ln K_i + \sum_{j=1}^{Nx} b_{i,j}\xi_j\right)\frac{\partial}{\partial \xi_j}\left(\ln K_i + \sum_{j=1}^{Nx} b_{i,j}\xi_j\right). \tag{A4}$$

Then, K_i does not depend on ξ_j, and $\frac{\partial}{\partial \xi_{j*}} = \sum_{j-1}^{Nx} b_{i,j}\xi_j = b_{i,j*}$. Thus, we have

$$\frac{\partial T_k}{\partial \xi_j} = \sum_{i=1}^{Nc} b_{i,k}b_{i,j}\exp\left(\ln K_i + \sum_{j=1}^{Nx} b_{i,j}\xi_j\right) = \sum_{i=1}^{Nc} b_{i,k}b_{i,j}[C_i]. \tag{A5}$$

More equations are included in the optimization when precipitates are present. These equations are solubility products, but they are treated as totals because they are part of the minimization as well:

$$\ln K_{S_l} = \sum_{j}^{Nx} bp_{l,j}\xi_j, \tag{A6}$$

where $l \in [1+Nx, Nx+Ncp]$. In this case, the derivative takes the following form:

$$\frac{\partial}{\partial \xi_j}\sum_{j=1}^{Nx} bp_{l,j}\xi_j = bp_{l,j}. \tag{A7}$$

In presence of precipitates we have additional unknowns $[Cp]$ that will be treated as regular concentrations $[C_i]$: the derivatives of Eq. A6 with respect to $[Cp]$ always equal zero because these equations only depend on ξ. Conversely, the derivatives of Eq. 9 with respect to $[Cp]$ are nonzero and can be easily computed by looking at the equation itself:

$$\left[\tilde{T}_j\right] - \sum_{i=1}^{Nc}\frac{b_{i,j}}{\gamma_i}K_i\prod_{k=1}^{Nx}\left(\gamma_j[X_k]\right)^{b_{i,k}} + \sum_{l=1}^{Ncp} bp_{l,j}[Cp_l] \quad j=1,\ldots,N_x. \tag{A8}$$

Notably, Eq. A1 contains a minus sign. If we consider the form of the linear system in the Newton Raphson algorithm, we can write the equation as follows:

$$-\mathbf{J}_n\Delta\mathbf{X}_n=\mathbf{Y}_n. \tag{A9}$$

Thus, the Jacobian matrix is computed in the code as the sole derivative of the computed totals, avoiding the minus sign.

Appendix B

The following tables provide stoichiometric coefficients, thermodynamic constants, and total concentrations for each test case presented in the work.

Table B1. Morel Table for Gallic Acid Test Case

	H^+	Al^{3+}	H_3L	$\log_{10}K$
H^+	1	0	0	0
Al^{3+}	0	1	0	0
H_3L	0	0	1	0
OH	−1	0	0	−14
H_2L	−1	0	1	−4.15
HL^2	−2	0	1	−12.59
L^3	−3	0	1	−23.67
$AlHL^+$	−2	1	1	−4.93
AlL	−3	1	1	−9.43
AlL_2^3	−6	1	2	−21.98
AlL_3^6	−9	1	3	−37.69
$Al_2(OH)_2(HL)_3L^2$	−8	2	3	−22.65
$Al_2(OH)_2(HL)_2L^3$	−9	2	3	−27.81
$Al_2(OH)_2(HL)L^4$	−10	2	3	−32.87
$Al_2(OH)_2L_3^5$	−11	2	3	−39.56
$Al_4L_3^{3+}$	−9	4	3	−20.25
$Al_3(OH)_4(H_2L)L^{4+}$	−5	3	1	−12.52
$[T_j]$	1.58E-06	1.00E-03	1.00E-03	

Table B2. Morel Table for Pyrite and Pyrite Mineral Test Cases

	H^+	O_2	Fe^{2+}	SO_4^2	$\log_{10}K$
H^+	1	0	0	0	0
OH	−1	0	0	0	−14
O_2	0	1	0	0	0
Fe^{2+}	0	0	1	0	0
$Fe(OH)_2$	−2	0	1	0	−20.6
$Fe(OH)_3^-$	−3	0	1	0	−31
$FeOH^+$	−1	0	1	0	−9.5
$FeSO_4$	0	0	1	1	2.2
Fe^{3+}	1	0.25	1	0	8.49
$Fe(OH)_2^+$	−1	0.25	1	0	2.82
$Fe(OH)_3^-$	−2	0.25	1	0	−3.51
$Fe(OH)_4^-$	−3	0.25	1	0	−13.11
$FeOH^{2+}$	0	0.25	1	0	6.3
$Fe_2(OH)_2^{4+}$	0	0.5	2	0	14.03
$Fe_3(OH)_4^{5+}$	−1	0.75	3	0	19.17
$Fe(SO)_2$	1	0.25	1	2	11.7
$FeSO_4^+$	1	0.25	1	1	10.4
SO_4^2	0	0	0	1	0
HSO_4^-	1	0	0	1	1.98
H_2SO_4	2	0	0	1	−1.02
SO_3^2	0	−0.5	0	1	−46.62
HSO_3^-	1	−0.5	0	1	−39.42
H_2SO_3	2	−0.5	0	1	−37.41
SO_2	2	−0.5	0	1	−37.56
$HS_2O_3^-$	3	−2	0	2	−132.52
$S_2O_3^2$	2	−2	0	2	−133.54
H_2S	2	−2	0	1	−131.33
HS^-	1	−2	0	1	−138.32

TABLE B2. *Continued*

	H^+	O_2	Fe^{2+}	SO_4^2	$\log_{10}K$
S^2	0	−2	0	1	−151.25
S_2^2	2	−3.5	0	2	−243.37
S_3^2	4	−5	0	3	−335.56
S_4^2	6	−6.5	0	4	−427.97
S_5^2	8	−8	0	5	−520.6
$S_2O_4^2$	2	−1.5	0	2	−118.46
$S_2O_5^2$	2	1	0	2	−83.65
$S_2O_6^2$	2	−0.5	0	2	−51.42
$S_2O_8^2$	2	0.5	0	2	−22.5
$S_3O_6^2$	4	−2	0	3	−146.1
$S_4O_6^2$	6	−0.35	0	4	−22.88
$S_5O_6^2$	8	−5	0	5	−332.54
$\boldsymbol{Fe(s)}$	−2	−0.5	1	0	−59.03
$\boldsymbol{FeS_2}$	2	−3.5	1	2	−217.40
$\boldsymbol{FeSO_4}$	0	0	1	1	−2.66
$[T_j]$	2.00E-03	−3.50E-03	1.00E-03	2.00E-03	

Table B3. Morel Table for Fe Cr and Fe Cr Min Test Cases

	H^+	$e-$	H_2CO_3	Fe^{2+}	CrO_4^2	K^+	SO_4^2	$\log_{10}K$
H^+	1	0	0	0	0	0	0	0
OH	−1	0	0	0	0	0	0	−14
O_2	−4	−4	0	0	0	0	0	−83.17
H_2	2	2	0	0	0	0	0	0
H_2O_2	−2	−2	0	0	0	0	0	−59
K^+	0	0	0	0	0	1	0	0
SO_4^2	0	0	0	0	0	0	1	0
KSO_4^-	0	0	0	0	0	1	1	0.5
H_2CO_3	0	0	1	0	0	0	0	0
HCO_3^-	−1	0	1	0	0	0	0	−6.3
CO_3^2	−2	0	1	0	0	0	0	−16.6
Fe^{2+}	0	0	0	1	0	0	0	0
$FeOH^+$	−1	0	0	1	0	0	0	−8.98
$FeOOH$	−3	0	0	1	0	0	0	−33.21
$Fe(OH)_{2(aq)}$	−2	0	0	1	0	0	0	−20.6
$Fe(OH)_3^-$	−3	0	0	1	0	0	0	−31
$Fe(OH)_4^2$	−4	0	0	1	0	0	0	−46
$FeHCO_3^+$	−1	0	1	1	0	0	0	−3.58
$FeSO_4$	0	0	0	1	0	0	1	1.2
Fe^{3+}	0	−1	0	1	0	0	0	−13.07
$FeOH^{2+}$	−1	−1	0	1	0	0	0	−15.2
$Fe(OH)_2^+$	−2	−1	0	1	0	0	0	−19.97
$Fe(OH)_4^-$	−4	−1	0	1	0	0	0	−34.62
$Fe_2(OH)_2^{4+}$	−2	−2	0	2	0	0	0	−29
$FeSO_4^+$	0	−1	0	1	0	0	1	−9.02
$Fe(SO_4)_2^-$	0	−1	0	1	0	0	2	−7.62
Cr^{2+}	8	4	0	0	1	0	0	68.09
$CrOH^+$	7	4	0	0	0	0	0	62.56
Cr^{3+}	8	3	0	0	1	0	0	74.98
$CrOH^{2+}$	7	3	0	0	1	0	0	71.17
$Cr(OH)_2^+$	6	3	0	0	1	0	0	64.95
$Cr(OH)_3^-$	5	3	0	0	1	0	0	80.9
$Cr(OH)_4^-$	4	3	0	0	1	0	0	47.5
CrO_4^2	0	0	0	0	1	0	0	0
$HCrO_4^-$	1	0	0	0	1	0	0	6.45
H_2CrO_4	2	0	0	0	1	0	0	6.3
$FeCrO_4^+$	0	−1	0	1	1	0	0	−19.31
$CrSO_4^+$	8	3	0	0	1	0	1	76.3
$KCrO_4^-$	0	0	0	0	1	1	0	0.57
$\boldsymbol{Fe(s)}$	0	2	0	1	0	0	0	13.12
$\boldsymbol{Cr(OH)_3(s)}$	5	3	0	0	1	0	0	66.1
$\boldsymbol{Cr_{(0,25)}Fe_{(0,75)}}$ OOH(s)	−1	0	0	0.25	0.75	0	0	2.93
$[T_j]$	0	0	10^{-6}	9.10^{-3}	3.10^{-3}	6.10^{-3}	9.10^{-3}	

Table B4. Morel Table for MoMaS Easy Test Case

	X_1	X_2	X_3	X_4	S	$\log_{10}K$
X_1	1	0	0	0	0	0
X_2	0	1	0	0	0	0
X_3	0	0	1	0	0	0
X_4	0	0	0	1	0	0
C_1	0	−1	0	0	0	−12
C_2	0	1	1	0	0	0
C_3	0	−1	0	1	0	0
C_4	0	−4	1	3	0	−1
C_5	0	4	3	1	0	35
S	0	0	0	0	1	0
CS_1	0	3	1	0	1	6
CS_2	0	−3	0	1	2	−1
$[T_j]$	0.3	0.3	0.3	2	10	

Table B5. Morel Table for MoMaS Hard Test Case

	X_1	X_2	X_3	X_4	X_5	S	$\log_{10}K$
X_1	1	0	0	0	0	0	0
X_2	0	1	0	0	0	0	0
X_3	0	0	1	0	0	0	0
X_4	0	0	0	1	0	0	0
X_5	0	0	0	0	1	0	0
C_1	0	−1	0	0	0	0	−12
C_2	0	1	1	0	0	0	0
C_3	0	−1	0	1	0	0	0
C_4	0	−4	1	3	0	0	−1
C_5	0	4	3	1	0	0	35
C_6	0	10	3	0	0	0	32
C_7	0	−8	0	2	0	0	−4
S	0	0	0	0	0	1	0
CS_1	0	3	1	0	0	1	6
CS_2	0	−3	0	1	0	2	−1
$\boldsymbol{Cp_1}$	0	3	1	0	0	0	10.9
$\boldsymbol{Cp_2}$	0	1	0	0	1	0	1.3
$[T_j]$	0.3	0.3	0.3	2	0.3	10	

Appendix C

In Table 6 are listed the $\log_{10}$ concentrations of the components that constitute the boundaries of the permitted intervals. When $X_1=H^+$, the component concentration ranges between 0 and −14 coherently with the pH domain. Otherwise, the minimum value is linked to Avogadro's constant.

Table C2. Limitations that are Imposed on the Computation of the Newton Raphson Increments and on the Activities of the Components

Maximum NR increment	$\max\lvert\Delta X_j\rvert$	5
Max concentration of $X_j=H^+$	$\max(X_j+\Delta X_j)=$	$9+\max(H^+)$
Min concentration of $X_j=H^+$	$\min(X_j+\Delta X_j)=$	$-9-\min(H^+)$
Max concentration of X_j	$\max(X_j+\Delta X_j)=$	$8\max(X)$

The values for the second component X_2 of the Fe Cr and Fe Cr Mineral test cases come from the following considerations regarding electric potential. H^+ is allowed to vary within the pH domain, i.e., from 0 to 14, while electrons e^- are allowed to generate a variation in electric potential from −1.2 to +1.2 V. The electron activity and electric potential are linked through the following relationships at 25°C:

$$\mathrm{Eh}=-\frac{\mathrm{RT}}{\mathrm{F}}\ln\{\mathrm{e}^-\}$$

$$\{\mathrm{e}^-\}=\exp\left(-\frac{\mathrm{F}}{\mathrm{RT}}\mathrm{Eh}\right)$$

$$\mathrm{Eh}=-\frac{8.314\times(273.15+25)}{96485,3399}\ln\{\mathrm{e}^-\}-2.569\times10^{-2}\times\ln\{\mathrm{e}^-\}.$$

For an electric potential of Eh −1.23 V, $\{e^-\}$ 1.6×10^{-21}. For Eh 1.23 V, $\{e^-\}$ $6.2\ 10^{20}$. Conversely, the values for the second component of the Pyrite and Pyrite Mineral tests come from the following reaction:

$$O_2+4H^++4e^-\rightleftarrows2H_2O.$$

The concentrations of H^+ and activity of O_2 are linked through the following relationships:

$$\mathrm{Eh}=\mathrm{E}^0+\frac{\mathrm{RT}}{4\mathrm{F}}\times\ln\left(\frac{\{O_2\}[H^+]^4}{\{H_2O\}}\right)=E^0+\frac{\mathrm{RT}}{4\mathrm{F}}\times\ln\left(\{O_2\}[H^+]^4\right)$$

$$(\mathrm{Eh}-E^0)\times\frac{4\mathrm{F}}{\mathrm{RT}}=\ln\left(\{O_2\}[H^+]^4\right)$$

$$\{O_2\}=[H^+]^4\exp\left[(\mathrm{Eh}-E^0)\times\frac{4\mathrm{F}}{\mathrm{RT}}\right].$$

For a pH of 14 and Eh 1.23 V, $\{O_2\}$ 1×10^{56}. For a pH 1 and Eh −1.23 V, $\{O_2\}$ 4.6×10^{-167}.

Table C1. Intervals for the Initial Guesses when Searching for Thermodynamic Equilibrium

		X_1	X_2	X_3	X_4	X_5	X_6	X_7	Cp_1	Cp_2	Cp_3
Gallic acid	min	−14	−23	−23							
	max	0	0	0							
Pyrite	min	−14	−164	−23	−23						
	max	0	54	0	0						
Pyrite M.	min	−14	−164	−23	−23				0	0	0
	max	0	54	0	0				0	0	0
MoMaS E.	min	−23	−23	−23	−23	−23					
	max	0	0	0	0	0					
MoMaS H.	min	−23	−23	−23	−23	−23	−23		0	0	
	max	0	0	0	0	0	0		0	0	
Fe Cr	min	−14	−21	−23	−23	−23	−23	−23			
	max	0	21	0	0	0	0	0			
Fe Cr Min.	min	−14	−21	−23	−23	−23	−23	−23	0	0	0
	max	0	21	0	0	0	0	0	0	0	0

Activities expressed in $\log_{10}$ scale for components X_j and linear scale for concentration of precipitates Cp_i

In Table 7 are listed all the limitations that are imposed on the computation of the Newton Raphson increments and on the activity of components, where $\max H\ln(1.0)$, $\min H^{+}\ln(10^{-14})$ and $\max X2$. No boundaries are imposed if the component is the activity of electrons. The only boundaries that are imposed on the concentrations of precipitates are the numbers that are allowed by the machine.

Manuscript received Feb. 23, 2016, and revision received Sep. 12, 2016.

Annexe 6. Implementation of Richardson extrapolation in an efficient adaptive time stepping method: applications to reactive transport and unsaturated flow in porous media

Transp Porous Med (2007) 69:123–138
DOI 10.1007/s11242-006-9090-3

ORIGINAL PAPER

Implementation of Richardson extrapolation in an efficient adaptive time stepping method: applications to reactive transport and unsaturated flow in porous media

Benjamin Belfort · Jérôme Carrayrou · François Lehmann

Received: 16 December 2005 / Accepted: 21 September 2006 / Published online: 30 January 2007

Abstract Environmental studies are commonly carried out through numerical simulations, which have to be accurate, reliable and efficient. When transient problems are considered, the validity of the solutions requires the calculation and management of the temporal discretization errors. This article describes an adaptive time stepping strategy based on the estimation of the local truncation error via the Richardson extrapolation technique. The time-marching scheme is mathematically based on this a posteriori error estimation that has to be gauged. General optimizations are also suggested making the control of both the temporal error and the evolution of the time step size very efficient. Furthermore, the algorithm connecting these methods is all the more interesting as it could be implemented in many computational codes using different numerical schemes. In the hydrogeochemical domain, this algorithm represents an interesting alternative to a fixed time step as shown by the various numerical tests involving reactive transport and unsaturated flow.

Keywords Richardson extrapolation · Adaptive time stepping · Reactive transport · Unsaturated flow

1 Introduction

Even if it can never replace experiments and field studies, modelling is of interest in many science and engineering applications for scientific understanding and/or technological management. In such an approach, ordinary- or partial-differential equations (ODE and PDE) are commonly used to develop mathematical models describing unsteady phenomena. The resolution of these equations through numerical

B. Belfort · J. Carrayrou · F. Lehmann (✉)
Institut de Mécanique des Fluides et des Solides, UMR 7507 ULP-CNRS,
2 rue Boussingault, Strasbourg, 67000, France
e-mail: lehmann@imfs.u-strasbg.fr

approximation leads to temporal, and often spatial discretizations, that invariably introduce numerical errors.

Since analytical solutions of the problem are often not known, the error may not be determined exactly and must be approximated in some way. The classical a priori theory provides or tries to determine a bound on the discretization error before the computation of the solution. It can become a challenge to obtain this bound with a sufficient accuracy. In fact, this depends on the convergence rate and on the derivatives of the function, which are both related to the particularities of the numerical scheme and the problem. Nevertheless a priori methods have been developed in various numerical schemes implemented in problems dealing with porous media. Recent applications are available (Schneid et al. 2004; Bause and Merz 2005; Sun and Wheeler 2005). A posteriori techniques give an estimation of the error, as a function of the results just obtained. Either the error estimation is in accordance with the numerical scheme (Babuska and Rheinboldt 1978; Zienkiewicz and Zhu 1992; Bank and Smith 1993), or it can be based on extrapolation techniques. In the last category can be found order- or grid-extrapolation error estimators. Predictor corrector approach or embedded Runge Kutta formulas are classical and efficient methods based on order extrapolation. However, it can be difficult to programme these methods, which need very specific modifications depending on the complexity of the problem. Perhaps less adapted for specific problems with non-linearities or complex geometries, an interesting aspect of the extrapolation-based error estimator is the possiblility of its implementation in a wide variety of calculation codes. From our point of view, this advantage justifies the attention we will confer to the Richardson extrapolation method. Many papers deal with the Richardson extrapolation, which is also referred to as the doubling method or h^2-/h^4-extrapolation. Hence, considerations for using the doubling method can be for instance, the mathematical convergence aspects (Ayati and Dupont 2004,Aïd 1999), the increase of accuracy order (Abbasian and Carey 1997) or the applicability to both time and spatial grids (Richards 1997).

Focusing on the temporal approximation, a natural connection for error estimation is its management through an optimal adaptive step size strategy. For the grid adaptation process, a priori methods relate the truncation error to the step size evolution coefficient. Nonetheless this relation is not necessary mathematically based i.e. heuristic parameters are included to increase or decrease the time step. Otherwize, the error estimation and the time-marching scheme are simply dissociated. Actually, an adaptive time stepping algorithm can also be developed by means of heuristic methods. This means for instance that the number of iterations achieved by an iterative solver can be used to define the next time step size. This kind of procedure requires a good appreciation on both the physical problem solved and the numerical method used.

In the view of temporal discretization error that invariably arises in numerical approximations, control of the temporal error and optimization of the time step are of great importance. Consequently, our main contribution consists in showing the efficiency of the Richardson extrapolation when combined with an a priori mathematical-based time stepping strategy, which really differs from fixed or heuristic control. The principle and demonstration of the Richardson extrapolation can be found in Richardson (1910; 1927), Shampine (1985) or in Hairer et al. (2000). The main results of this grid-extrapolation technique are depicted at the beginning of the article to present the time stepping algorithm we focus on. Then, several techniques dealing with the estimated error, the choice of the initial time step, or the initialization in an iterative process are proposed in the part entitled "optimization of the method".

The general formulation of the algorithm allows treatment of a large variety of non-linear physical processes with very different time scales. They also involve rather different mathematical models and specific numerical solutions. Consequently, the proposed time-marching scheme and optimizations have been incorporated in different codes describing reactive transport and unsaturated flow in porous media. Several test cases are performed to illustrate the interest of monitoring both the local error and the time step size.

2 Presentation of the method

The main idea of the Richardson extrapolation is to solve the same problem first in one large time step and second in two half time steps. These approximations are used to estimate the local truncation error. This estimation can be used to define the length of the next time step and therefore allows the development of an efficient automatic and adaptive time stepping algorithm.

2.1 Extrapolation

Let Eq. 1 be the general form of an ODE, a system of ODE, a PDE or a system of PDE.

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f\left(t, y\left(t\right)\right). \tag{1}$$

Assuming that the numerical method used is of p order in time, the difference between the exact value of the variable, y_{Ex}^{n+1}, and the approximate one obtained in a single step, $\tilde{y}^{n+1,*}$, at $t = n + 1$, is the error given by the approximation (Shampine 1985; Hairer et al. 2000):

$$y_{\mathrm{Ex}}^{n+1} - \tilde{y}^{n+1,*} = A\Delta t^{p+1} + O\left(\Delta t^{p+2}\right), \tag{2}$$

where A depends on the size of the derivatives of the solution in the interval.

For a sufficiently smooth function f, the local error of the two steps viewed as a single step can be expressed as follow (Shampine 1985; Hairer et al. 2000):

$$y_{\mathrm{Ex}}^{n+1} - \tilde{y}^{n+1,**} = 2A\left(\frac{\Delta t}{2}\right)^{p+1} + O\left(\Delta t^{p+2}\right), \tag{3}$$

where $\tilde{y}^{n+1,**}$ is the variable obtained in two steps.

Hence, neglecting terms of order higher than $p + 1$ and combining Eqs. 2 and 3 gives:

$$A = \frac{2^p}{\Delta t^{p+1}} \frac{\tilde{y}^{n+1,**} - \tilde{y}^{n+1,*}}{2^p - 1}. \tag{4}$$

An extrapolated solution, $y_{\mathrm{extrap}}^{n+1}$, of order $p + 1$ can be calculated as follow:

$$y_{\mathrm{extrap}}^{n+1} = \tilde{y}^{n+1,**} + \frac{\tilde{y}^{n+1,**} - \tilde{y}^{n+1,*}}{2^p - 1}. \tag{5}$$

2.2 Time step size adaptation

The error of this method corresponds to the difference between the exact value of the variable, y_{Ex}^{n+1}, and the approximate one:

$$\text{Err}_i\left(\Delta t\right) = \left| y_{\text{Ex},i}^{n+1} - y_{\text{extrap},i}^{n+1} \right|, \quad i = 1, \ldots, \text{NN}, \tag{6}$$

where NN refers to the dimension of the solution vector.

Since y_{extrap}^{n+1} is a local extrapolation of order $p+1$, the following inequality is proposed:

$$\text{Err}_i\left(\Delta t\right) \leq \left| y_{\text{Ex},i}^{n+1} - \tilde{y}_i^{n+1,**} \right|, \quad i = 1, \ldots, \text{NN}. \tag{7}$$

Due to the fact that the accuracy of the extrapolated solution is unknown, inequality (7) is assumed to be correct and an estimated error, $\text{Err}_{\text{est}}\left(\Delta t\right)$, is then calculated. Hence, Eqs. 3 and 4 can be combined and then inserted in expression (7) to define the estimated error, which has to be gauged using the following inequality:

$$\text{Err}_{\text{est},i}\left(\Delta t\right) = \left| \frac{\tilde{y}_i^{n+1,**} - \tilde{y}_i^{n+1,*}}{2^p - 1} \right| \leq \varepsilon_i = \varepsilon_{\text{a}} + \varepsilon_{\text{r}} \left| y_{\text{extrap},i}^{n+1} \right|, \quad i = 1, \ldots, \text{NN}. \tag{8}$$

In the previous equation, ε_i is the precision criterion we want to respect by adjusting the time step length. This mixed type of error control includes an absolute, ε_{a}, and a relative, ε_{r}, truncation error tolerance.

Assuming that a calculation is performed with the time step $\Delta t_{\text{current}}$, an estimation of the error for this time step, $\text{Err}_{\text{est}}\left(\Delta t_{\text{current}}\right)$, is calculated. This estimation can be smaller or greater than ε_i. Independently of the result obtained in Eq. 8, a new time step Δt_{new} must be calculated, either to estimate the variable y at time $n+2$ or to improve the accuracy at time $n+1$. Equation 4 and the definition of the estimated error give:

$$A_i = \frac{2^p}{\Delta t_{\text{current}}^{p+1}} \text{Err}_{\text{est},i}\left(\Delta t_{\text{current}}\right), \quad i = 1, \ldots, \text{NN}. \tag{9}$$

Assuming that A is unchanged, i.e. f is considered (sometimes by extension) as smooth, the respect of the criterion (8) implies that the new time step should fulfil the Eq. 10:

$$A_i = \frac{2^p}{\Delta t_{\text{new}}^{p+1}} \varepsilon_i, \quad i = 1, \ldots, \text{NN}. \tag{10}$$

Simplifying A in both Eqs. 9 and 10 provides an estimation of the new time step:

$$\Delta t_{\text{new}} = \sqrt[p+1]{\min_{i=1,\ldots,\text{NN}} \left| \frac{\varepsilon_i}{\text{Err}_{\text{est},i}\left(\Delta t_{\text{current}}\right)} \right|} \Delta t_{\text{current}}. \tag{11}$$

If the current time step is sufficiently small, then the estimated error is smaller than the truncation error tolerance, so that the factor multiplying the current time step is greater than one and the new time step consequently increases. Otherwise the calculation of y_{extrap}^{n+1} is then rejected and should be repeated with a smaller time step.

An algorithm is also implemented to avoid large changes in the time step evolution around output times (Kavetski et al. 2001).

2.3 Optimization of the method

Some adjustments should be made to increase the efficiency of the method. They deal with the control of the time step size. Specifications due to the implementation of the Richardson extrapolation for the resolution of non-linear system or for the initialization strategy are also reported.

2.3.1 Relative test and tolerance on the precision criterion

For many applications described with PDE or systems of ODE, the variable y is a vector in which component values can vary over several orders of magnitude. In this case, a strictly relative test ($\varepsilon_a = 0$) can be attractive and has been kept in the examples performed in the next section.

To avoid too many failed steps, a safety factor can be introduced to relax the time step size evolution (Hairer et al. 2000). An other possibility consists in relaxing the truncation error test with a factor Tol:

$$\mathrm{Err}_{\mathrm{est},i}(\Delta t) \leq \varepsilon_i \times \mathrm{Tol}, \quad i = 1, \ldots, \mathrm{NN}, \tag{12}$$

where Tol refers to a tolerance on the precision criterion, which lies between 2 and 10 (Tol = 5 in this article).

Practically, this tolerance means that the calculation can be accepted even if the estimated error is Tol times larger than the precision criterion. The time step size control formula (11) does not change. Therefore, it may be noticed that even if a calculation is accepted with an estimated error higher than ε due to the tolerance, the next time step size is determined to give an estimated error equal to ε. This leads then to a reduction of the time step size.

If the computing time of one time step is great, for PDE over a large domain for instance, this procedure avoids too many failed steps, which are CPU time consuming.

2.3.2 Selection of the first time step

The choice of the first time step is the most empirical decision for such a method. It can be selected from previous experiences in computation of similar problems or from other considerations such as stability conditions of the numerical method (Courant or Péclet number for example in the case of PDE).

A proposition for efficiently choosing the first time step is developed in the following part. Similar methods can be found in Hairer et al. (2000). Using a Taylor's expansion in the function f makes it possible to express A as:

$$A = \frac{\partial^p f}{\partial t^p}. \tag{13}$$

Equation 10 is supposed to be valid for the first time step and assuming that the derivative of Eq. 13 and can be evaluated, the first time step is given by

$$\Delta t_{\mathrm{first}} = \sqrt[p+1]{2p \min_{i=1,\ldots,\mathrm{NN}} \left| \frac{\varepsilon_a + \varepsilon_r \left| y_{i,t=0} \right|}{\left. \frac{\mathrm{d}^p f}{\mathrm{d}t^p} \right|_{y_{i,t=0}}} \right|}. \tag{14}$$

For high-order methods (fourth order Runge–Kutta for example), it seems that the best way to estimate the p derivative of f is to do this analytically. Nevertheless, if the p derivative is not known, we propose the following empirical relation to calculate the first time step length:

$$\Delta t_{\text{first}} = \left(2^p - 1\right) \cdot \sqrt{\min_{i=1,\ldots,\text{NN}} \left| \frac{\varepsilon_{\text{a}} + \varepsilon_{\text{r}} \left| y_{i,t=0} \right|}{f\left(t, y_i\right)_{t=\Delta\widetilde{t}} - f\left(t, y_i\right)_{t=0}} \right| \times 2^p \Delta\widetilde{t}}, \tag{15}$$

where $\Delta\widetilde{t}$ has to be chosen sufficiently small depending on the characteristic time of the simulation and the precision criterion.

In the case of the first order method, the derivative is also easier to evaluate numerically with Eq. 15.

2.3.3 Implementation for non-linear ODE or PDE

The algorithm based on Richardson extrapolation can also be used for non-linear problem. The linearization with iterative methods requires an initial guess, which can be estimated with a predictor technique for the first big step and trapezoidal rules for the two half steps.

Difficulties can be observed when secondary variables or mass balance have to be calculated with the extrapolated solution, which does not necessary respect the convergence criterion checked by $\tilde{y}^{n+1/2}$, $\tilde{y}^{n+1,*}$ and $\tilde{y}^{n+1,**}$. The examples developed in the next section provide interesting illustrations of this kind of problem and the specific ways to solve them. The first idea consists in solving again the system with the extrapolated solution as an initial guess and especially with a higher order method. If a first order Euler discretization is initially used, it means that a Crank–Nicolson scheme should be implemented. This strategy maintains the accuracy's order of y^{n+1}_{extrap}. A technique, that has also been tested, is a generalization of the extrapolation for all the variables used.

3 Examples

Two examples are solved with the optimized algorithm. They deal successively with reactive transport and unsaturated flow. Reactive transport modelling leads to a differential and algebraic system representing the coupled solution of chemical reaction and solute transport equations. On the one hand, the advective-dispersive solute transport equation behaves as hyperbolic when transport is advection dominated, or parabolic when dispersion dominates. On the other hand, instantaneous equilibrium chemistry is described by a non-linear algebraic system. The first test case includes field observations published by Valocchi et al. (1981) and serves subsequently as a benchmark problem for testing reactive transport codes. It deals with an advection dominated transport associated with non-linear cation-exchange reactions. Besides, unsaturated flow is described with a highly non-linear parabolic equation, which is really challenging to solve when sharp infiltration fronts are simulated. Therefore, the classical benchmark scenario described by Celia et al. (1990) is used to check the robustness of the numerical method.

In each part, after a short presentation of the method developed to treat the problem, we specify the model traditionally used in the context, the implementation of the algorithm and the test case with its conclusions.

The ability of the proposed time-marching scheme to control temporal errors is assessed using two kinds of error measurements. The first one, which could be called cumulated relative error measure versus the reference solution ($\mathrm{CREM_{ref}}$), collects the relative error produced at each time step until the end of simulation:

$$\mathrm{CREM_{ref}} = \sum_{t=1}^{\text{Nb time steps}} \left| \frac{y_{\mathrm{NN},t} - y_{\mathrm{ref,NN},t}}{y_{\mathrm{ref,NN},t}} \right|, \tag{16}$$

where $y_{\mathrm{ref,NN},t}$ refers to the reference solution at node NN and time t. It corresponds to elution curve.

The differences between the profiles of the computational results and the reference solution can be integrated along the spatially discretized domain at any observation time where the reference is known. This relative error measure ($\mathrm{REM}_{\mathrm{ref}}^{n+1}$) is defined at time $n+1$ with:

$$\mathrm{REM}_{\mathrm{ref}}^{n+1} = \sum_{i=1}^{\mathrm{NN}} \left| \frac{y_i^{n+1} - y_{\mathrm{ref},i}^{n+1}}{y_{\mathrm{ref},i}^{n+1}} \right|, \tag{17}$$

where $y_{\mathrm{ref},i}^{n+1}$ is the reference solution at time $n+1$ and node i.

3.1 Reactive transport with operator splitting

The Richardson extrapolation is usually applied to steady state problems to estimate the spatial truncation error and to adapt the grid size. Also, it has been carried out for advection diffusion problems (Natividad and Stynes 2003) and for advection-diffusion-reaction problems describing laminar flames (Claramunt et al. 2004). Richardson extrapolation has also been used to increase the temporal accuracy of reaction-diffusion equations solved by a global approach (Liao et al. 2002). Nevertheless, these authors did not used the ability of Richardson extrapolation to provide adaptive time stepping. Therefore, the algorithm combining the time step selection and the error estimated with the Richardson extrapolation could be originally developed in the context of transient flow for hydrogeochemical calculations. The control of the error is all the more important because the standard non-iterative operator splitting approach used in this work can introduce some temporal error due to the discretization (Carrayrou et al. 2003).

3.1.1 Presentation of the model

The reactive transport equation for porous media is written, under the instantaneous equilibrium assumption (Rubin 1983; Steefel and McQuarrie 1996):

$$\frac{\partial \left(\omega T_{\mathrm{d}} + \rho_{\mathrm{s}} T_{\mathrm{f}}\right)}{\partial t} = \nabla \cdot \left[D \cdot \nabla \left(T_{\mathrm{d}}\right)\right] - U \cdot \nabla \left(T_{\mathrm{d}}\right), \tag{18}$$

where T_{d} is the total mobile (dissolved) component concentration, T_{f} is the total immobile (fixed) component concentration, ω is the porosity of the media, ρ_{s} is the density of the solid matrix, U is the Darcy velocity and D is the dispersion coefficient.

For a given total mobile plus immobile concentration of components, solving the algebraic system describing instantaneous equilibrium gives the concentration of each component each species and then the distribution of the component between the mobile and immobile phases. This is summarized as:

$$\begin{cases} T_{\mathrm{d}} = f_{\mathrm{d}}\left(\omega T_{\mathrm{d}} + \rho_{\mathrm{s}} T_{\mathrm{f}}\right) \\ T_{\mathrm{f}} = f_{\mathrm{f}}\left(\omega T_{\mathrm{d}} + \rho_{\mathrm{s}} T_{\mathrm{f}}\right) \end{cases}, \tag{19}$$

where f_{d}, respectively f_{f}, represent the non-linear algebraic systems describing chemistry in aqueous and solid phases, respectively.

Combining the transport Eq. 18 and the chemical laws (19) leads to a non-linear differential algebraic system. One of the simplest way to solve this system is to split it between both transport and chemistry operator (Yeh and Tripathi 1989; Carrayrou et al. 2004). The standard non-iterative scheme has been used in this work. Since the error introduced by this operator splitting approach depends on time discretization (Carrayrou et al. 2003), the Richardson extrapolation and the associated adaptive time-marching scheme provide an interesting means to control the error.

In this work, the transport operator includes an implicit first order time discretization and a finite volume method. The transport operator is first solved for all components assuming they are not reactive (20):

$$\omega \frac{T_{\mathrm{d}}^{n+1,T} - T_{\mathrm{d}}^{n}}{\Delta t} = \nabla \cdot \left[D \cdot \nabla\left(T_{\mathrm{d}}\right)\right] - U \cdot \nabla\left(T_{\mathrm{d}}\right). \tag{20}$$

This leads to an intermediate solution $T_{\mathrm{d}}^{n+1,T}$, which is used as an initial condition for the chemistry operator:

$$\begin{cases} T_{\mathrm{d}}^{n+1} = f_{\mathrm{d}}\left(\omega T_{\mathrm{d}}^{n+1,T} + \rho_{\mathrm{s}} T_{\mathrm{f}}^{n}\right) \\ T_{\mathrm{f}}^{n+1} = f_{\mathrm{f}}\left(\omega T_{\mathrm{d}}^{n+1,T} + \rho_{\mathrm{s}} T_{\mathrm{f}}^{n}\right) \end{cases} \tag{21}$$

The chemistry operator is solved at each grid point using a combined algorithm associating the definition of the chemical allowed intervals, a preconditioning by positive continuous fraction method and a Newton–Raphson method (Carrayrou et al. 2002). The solutions of the chemistry operator T_{d}^{n+1} and T_{f}^{n+1} are the solutions of the standard non-iterative scheme at the time step $n+1$. It is well known that this scheme increases the numerical diffusion, but is also useful to solve the convergence problems related to iterative schemes.

3.1.2 Implementation of the time stepping method with the Richardson extrapolation

As presented previously, the overall system (20) and (21) is solved three times for each time step leading to $T_{\mathrm{d}}^{n+1,*}$, $T_{\mathrm{d}}^{n+1,**}$, $T_{\mathrm{f}}^{n+1,*}$ and $T_{\mathrm{f}}^{n+1,**}$. The extrapolation (5) is done for both variables T_{d}^{n+1} and T_{f}^{n+1} at each cell of the space discretization and for each chemical component. Estimated errors are calculated for both the total dissolved and the total fixed concentrations for each component and at each cell of the mesh. All of them have to verify Eq. 12 and the smallest time step coming from Eq. 11 is used. Therefore, the required precision for all the variables is ensured at the current time and should be at the next step.

Table 1 Physico-chemical parameters for the reactive transport test-case

	Cl^-	Na^+	Ca^{2+}	Mg^{2+}
Initial ($mg\,l^{-1}$)	5,700	1,990	444	436
Injected ($mg\,l^{-1}$)	320	216	85	12
Cation exchange capacity ($meq\,g^{-1}$)		0.1		
Bulk density ($g\,l^{-1}$)		1,875		
Porosity		0.25		
Dispersivity (m)		2.96		
Length of the domain (m)		16		
Darcy velocity ($m\,s^{-1}$)		0.2525		
Spatial discretization (m)		0.1		

Since the extrapolated total concentrations calculated with Eq. 5 do not respect the chemical equilibrium condition, the instantaneous equilibrium system (21) is solved one more time after the extrapolated solution (22) is known.

$$\begin{aligned} T_{\mathrm{d}_i}^{n+1} &= 2T_{\mathrm{d}_i}^{n+1,**} - T_{\mathrm{d}_i}^{n+1,*}, \\ T_{\mathrm{f}_i}^{n+1} &= 2T_{\mathrm{f}_i}^{n+1,**} - T_{\mathrm{f}_i}^{n+1,*}. \end{aligned} \tag{22}$$

3.1.3 Test case and discussion

An experiment described by Valocchi et al. (1981) has been tested numerically. It presents the injection of water into the aquifer at the Palo Alto Baylands Field site. The chemical phenomena concern ion exchange, described by Eq. 23. Physico-chemical conditions of the test case are given in Table 1. Cl^- appears in this table to ensure electroneutrality.

$$\begin{aligned} 2 \cdot (\equiv \mathrm{S-Na}) + \mathrm{Ca}^{2+} &\rightleftarrows (\equiv \mathrm{S_2-Ca}) + 2 \cdot \mathrm{Na}^{+} \text{ with } \mathrm{K_{NaCa}} = 1.7\mathrm{eq/L} \\ 2 \cdot (\equiv \mathrm{S-Na}) + \mathrm{Mg}^{2+} &\rightleftarrows (\equiv \mathrm{S_2-Mg}) + 2 \cdot \mathrm{Na}^{+} \text{ with } \mathrm{K_{NaMg}} = 3.0\mathrm{eq/L} \end{aligned} \tag{23}$$

In Fig. 1, elution curves for calcium and magnesium given by the adaptive time step with and without extrapolation are compared. A reference solution is obtained with a very small precision criterion and is validated by comparison with experimental data given by Valocchi et al. (1981). This figure illustrates clearly the increase of precision induced by the extrapolation explained in Eq. 5. The extrapolated elution curve is closer to the reference solution than the non-extrapolated one.

Using the extrapolated solution (5) or (22) leads to a more accurate solution at the cost of one more solution of the instantaneous equilibrium. Although this involves additional computation, the extrapolation with adaptive time stepping presented in this work is very efficient, as can be seen in Fig. 2. $\mathrm{CREM_{ref}}$ has been calculated for each component and the maximum value has been plotted. This figure shows that, as expected from the theory, a fixed time step and an adaptive time step without extrapolation leads to a first order relation between precision and CPU time. On the other hand, the combination of extrapolation and an adaptive time step gives a second order relation between precision and CPU time.

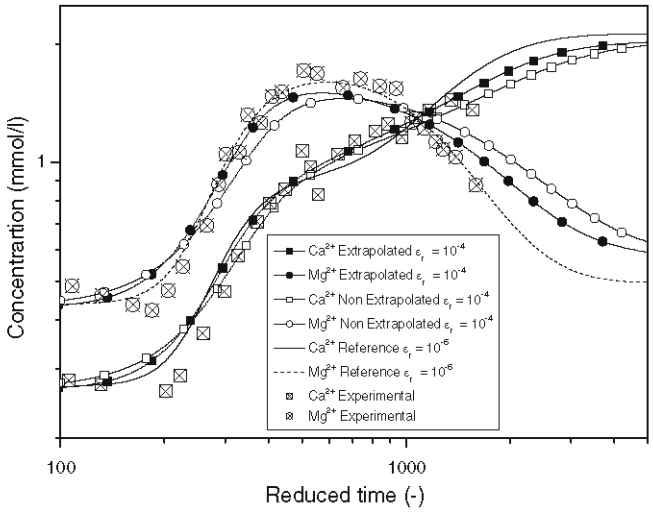


Fig. 1 Reactive transport test case: comparison of the elution curves

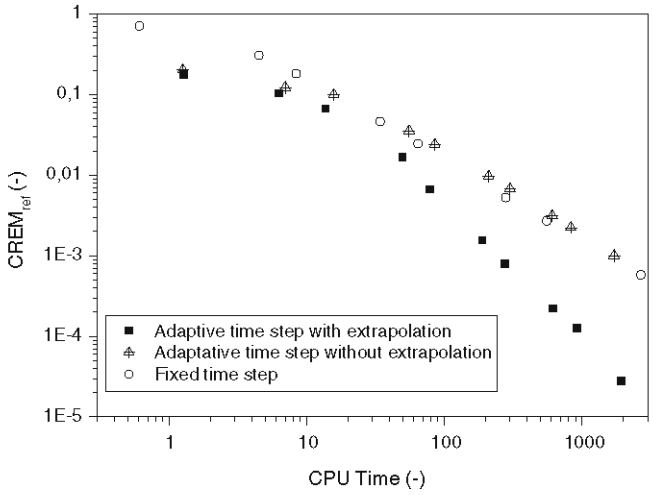


Fig. 2 Reactive transport test case: evolution of the Cumulated Relative error measure ($CREM_{ref}$) versus the required CPU time

3.2 Unsaturated flow

The Richardson extrapolation has been studied for groundwater flow applications. Guarracino et al. (2004) used the pressure head form of Richards' equation and associated the extrapolation with a Crank–Nicolson scheme to reach a third order accurate temporal scheme. The authors did not insert a time-marching scheme and verified principally the accuracy and the mass conservation properties. Besides, Basombrio et al. (2006) developed a competitive non-iterative algorithm combining Crank–Nicolson method, Richardson extrapolation and a single Newton's iteration. However, the amplification or reduction factor for the time step is quite heuristic.

3.2.1 Presentation of the model

The last example deals with the infiltration of water through an initially dry porous media. The mathematical model used to describe this physical problem is given by Eqs. 24 and 25.

Darcy–Buckingham's law defines the water flux in the domain:

$$q = -K(h).\nabla(h - z), \tag{24}$$

where q is the macroscopic fluid flux density, K is the hydraulic conductivity, h is the pressure head and z is the depth, taken to be positive downwards.

The mixed form of Richards' equation represents the mass conservation of water:

$$\frac{\partial\theta(h)}{\partial t} + \nabla \cdot q = f_{\mathrm{v}}, \tag{25}$$

where θ is the volumetric water content, t is time, f_{v} is a source/sink term, and q is the water flux previously defined.

To complete this description, the interdependencies of the pressure head, the hydraulic conductivity and the water content must be characterized using constitutive relations. The standard Mualem–van Genuchten (1980) is used:

$$\begin{aligned} S_{\mathrm{e}}(h) &= \frac{\theta(h) - \theta_{\mathrm{r}}}{\theta_{\mathrm{s}} - \theta_{\mathrm{r}}} = \begin{cases} \frac{1}{(1+(\alpha|h|)^n)^{1-(1/n)}} & h < 0 \\ 1 & h \geq 0 \end{cases} \\ K(S_{\mathrm{e}}) &= K_{\mathrm{s}} S_{\mathrm{e}}^{1/2} \left[1 - \left(1 - S_{\mathrm{e}}^{(n/(n-1))}\right)^{(1-(1/n))}\right]^2, \end{aligned} \tag{26}$$

where θ_{s} and θ_{r} are the saturated and residual volumetric water contents, respectively, α is a parameter related to the mean pore size and n a parameter reflecting the uniformity of the pore-size distribution ($n > 1$).

The numerical technique implemented is a traditional finite volume method for the spatial discretization and a backward Euler scheme for the temporal approximation. The interblock conductivities, which appear for the calculation of the flux between adjacent cells of the mesh, are calculated either with a geometric or an arithmetic mean. Due to the non-linearities of the constitutive relationships, we have to solve non-linear partial differential equations. The discretized system of PDE is linearized using the modified Picard (or fixed-point) method (Lehmann and Ackerer 1998). Iterations proceed until the mixed absolute-relative convergence test is satisfied:

$$\left|h_i^{n+1,k+1} - h_i^{n+1,k}\right| \leq \tau_{\mathrm{r}} \left|h_i^{n+1,k}\right| + \tau_{\mathrm{a}}, \quad i = 1, \ldots, \mathrm{NN}, \tag{27}$$

where k denotes the iteration number. τ_{a} and τ_{r} refer to absolute and relative convergence criteria. They are hundred times smaller than the corresponding criteria on the truncation error tolerance.

3.2.2 Implementation of the time stepping method with the Richarsdon extrapolation

After each time step, the pressure head and the water content are updated with the Richardson extrapolation:

$$\begin{aligned} h^{n+1} &= 2h^{n+1,**} - h^{n+1,*}, \\ \theta^{n+1} &= 2\theta^{n+1,**} - \theta^{n+1,*}. \end{aligned} \tag{28}$$

Table 2 Initial, boundary conditions and parameters values for the unsaturated flow test case

Parameters	Value
Material and/or	Sand
Reference	Celia et al. (1990)
$\theta_r(-)$	0.102
$\theta_s(-)$	0.368
α (cm^{-1})	0.0335
n (–)	2
K_s ($cm\,s^{-1}$)	9.22×10^{-3}
Initial conditions	
$h(z, t = 0)$ (cm)	−1,000
Boundary conditions	
$h(z = 0\ cm, t)$(cm)	−75
$h(z = 100\ cm, t)$(cm)	−1,000

The temporal accuracy of the scheme has been considered as an important criterion. However, the ability of the code to conserve good global mass balance is also essential. The extrapolated solutions presented in (28) have no real physical meaning. To avoid a large mass balance error, we suggested in a previous section to again solve the system with the extrapolated solution at each time step with a higher order numerical scheme. Another technique consists of extrapolating the flux.

3.2.3 Test case and discussion

We simulate a sharp infiltration front in a homogeneous dry porous media as proposed by Celia et al. (1990). The initial, boundary conditions and the relevant material properties are summarized in Table 2.

Figure 3 displays pressure head profiles after a half day of infiltration. The dense grid solution provided by Celia et al. (1990) has also been plotted to show the convergence of the method when the nodal spacing decreases. The interblock conductivity is averaged using the arithmetic mean. Figure 3 illustrates the interest of the extrapolation compared to fixed time step, time stepping scheme without extrapolation, or a heuristic time-marching scheme based on the behaviour of the non-linear iteration.

To investigate temporal aspects of the Richardson extrapolation in unsaturated water movement, a surrogate "reference" solution is evaluated numerically using the adaptive scheme with a relative error tolerance of $\varepsilon_r = 10^{-8}$ and a convergence criterion of $\tau_r = 10^{-10}$. An identical fixed-grid with a nodal spacing of 1 cm and a geometric interblock conductivity are used for all simulations thus making it possible to neglect spatial errors and to focus only on the temporal errors.

The proposed adaptive time stepping method allows the control of the temporal error with the relative tolerance criterion ε_r. An improvement of the precision coincides with the automatic decrease of the step size by the algorithm as depicted in Fig. 4. It shows a very classical evolution of the step size.

We observe that reducing the relative precision criterion by a factor of one hundred leads to a decrease of ten times the mean step size. In fact, the mean length of the time step reaches 450 s for the worst precision considered and just above 1 s for the largest.

To analyse the efficiency of the method, the relative error has been plotted as a function of the CPU time. Figure 5 is hence obtained by adjusting the criterion ε_r for the adaptive scheme or varying the time step size for the fixed step method. As

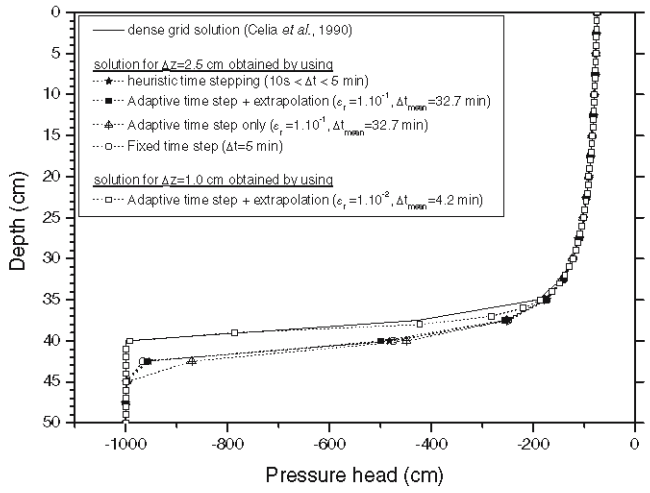


Fig. 3 Unsaturated flow test case: pressure head profiles after 12 h of infiltration

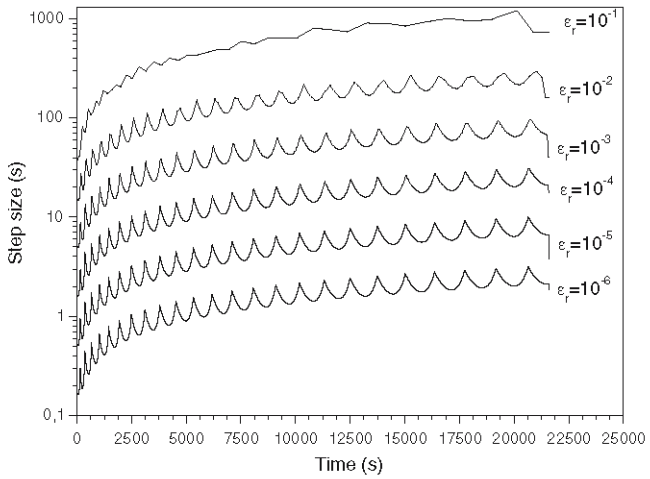


Fig. 4 Unsaturated flow test case: evolution of the time step size versus time for the scheme with extrapolation

shown in the previous example, Fig. 5 clearly illustrates that the algorithm using the Richardson extrapolation leads faster to a higher accuracy. Hence, the adaptive time stepping method becomes competitive when associated to the extrapolation.

It is all the more interesting because the mass balance can be correctly managed when some precautions are taken into account. With a constant nodal spacing, the formula commonly used to calculate the global mass balance is (Celia et al. 1990):

$$\mathrm{GMB}(\%) = \left| \frac{\left[\frac{1}{2} \times \left(\theta_1^{n+1} - \theta_1^0 \right) + \sum_{i=2}^{\mathrm{Ne}} \left(\theta_i^{n+1} - \theta_i^0 \right) + \frac{1}{2} \times \left(\theta_{\mathrm{Nn}}^{n+1} - \theta_{\mathrm{Nn}}^0 \right) \right] \times \Delta x}{\sum_{j=\mathrm{time_{init}}}^{\mathrm{time_{print}}} \left(q_1^j - q_{\mathrm{Nn}}^j \right) \times \Delta t^j} - 1 \right| \times 100 \quad (29)$$

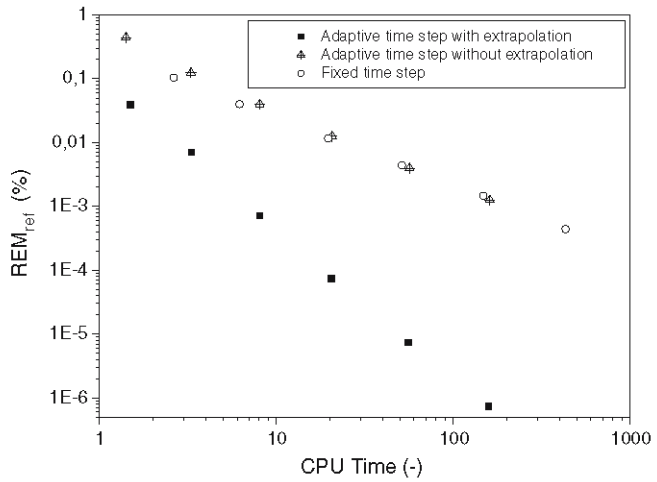


Fig. 5 Unsaturated flow test case: evolution of the relative error versus the required CPU time after 6 h of infiltration

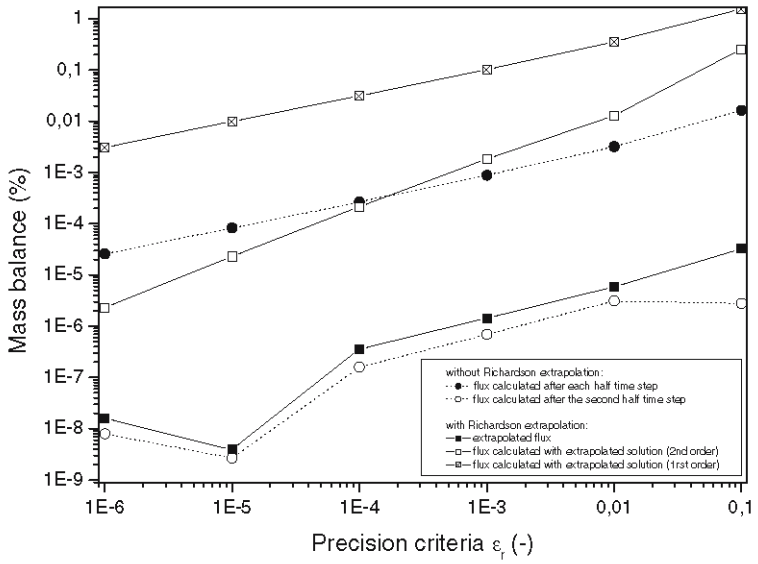


Fig. 6 Unsaturated flow test case: representation of the mass balance function of the relative precision criteria: comparisons of different techniques for the flux approximation

The fluxes that appear in the previous equation can be estimated using a variety of means. If the extrapolation is used at each time step for the error estimation and the variable adaptation, the flux can be calculated through a first (totally implicit formulation) or a second (Crank–Nicolson formulation) order approximation. Nevertheless, Fig. 6 shows that the best technique is to also extrapolate the flux. If the variables obtained in two time steps are retained, Fig. 6 also illustrates that the flux cannot be viewed as a general flux on this period calculated with the last pressure. This must be calculated after each half time step.

4 Conclusion

After a brief presentation of the Richardson extrapolation, this article has described a general way of taking into account the truncation error for an efficient management of the step size evolution. The automatic time-marching scheme is mathematically based and the user must only define the accepted tolerance on the temporal discretization error. Another important aspect of this work deals with optimization strategies to estimate the first time step, to implement the algorithm in a non-linear system, or to introduce flexibility in the time evolution i.e. to avoid too many rejected time steps. The whole of our approach was developed in a context that could allow its application to diverse numerical fields. This algorithm can easily take into account specificities of a given problem.

In fact, all our propositions have been implemented in rather different codes that model kinetic chemistry, reactive transport or unsaturated flow. The global formulation of the algorithm allows treatment of notably different mathematical models. The proposed method is an efficient way of adapting the time step size and of estimating the error for many problems frequently encountered in porous media.

The use of extrapolation technique appears advantageous. First, the examples show that the accuracy has been improved. For a given error calculated with a reference solution, the extrapolation of the variables yields a decrease in computation time compared to both a fixed time step or an adaptive evolution without extrapolation. Second, although extrapolation may not always have physical meaning, it can still conserve properties as illustrated in our example mass balance calculation.

Future research could deal with a comparison of different time- marching schemes, a spatial adaptation coupled with the time stepping strategy, or a separate time stepping procedure for the transport and reaction operators involved in the splitting method.

Acknowledgements The authors sincerely thank the seven anonymous reviewers and Paul Montgomery (CR-HDR CNRS Strasbourg) for their suggestions for improvements.

References

Abbasian, R.O., Carey, G.F.: A note on Richardson extrapolation as an error estimator for non-linear reaction-diffusion problems. Commun. Numer. Methods Eng. **13**(7), 533–540 (1997)

Aïd, R.: Richardson estimator with variable step-size. C. R. Acad. Sci. Paris **329**(I), 833–837 (1999)

Ayati, B.P., Dupont, T.F.: Convergence of a step-doubling galerkin method for parabolic problems. Math. Comput. **74**(251), 1053–1065 (2004)

Babuska, I., Rheinboldt, W.C.: A posteriori error estimates for finite element method. Int. J. Numer. Methods Eng. **12**(10), 1597–1615 (1978)

Bank, R.E., Smith, R.K.: A posteriori error estimates based on hierarchical bases. SIAM J. Numer. Anal. **30**, 921–932 (1993)

Basombrio, F.G., Guarracino, L., Vénere, M.J.: A non-iterative algorithm based on Richardson' extrapolation. Application to groundwater flow modelling. Int. J. Numer. Methods Eng. **65**(7), 1088–1112 (2006)

Bause, M., Merz, W.: Higher order regularity and approximation of solutions to the Monod biodegradation model. Appl. Numer. Math. **55**(2), 154–172 (2005)

Carrayrou, J., Mosé, R., Behra, Ph.: New efficient algorithm for solving thermodynamic chemistry. AIChE. J. **48**(4), 894–904 (2002)

Carrayrou, J., Mosé, R., Behra, Ph.: Modelling reactive transport in porous media: iterative scheme and combination of discontinuous and mixed-hybrid finite elements. C. R. Mec. **331**(3), 211–216 (2003)

Carrayrou, J., Mosé, R., Behra, Ph.: Efficiency of operator splitting procedures for solving reactive transport equation. J. Contam. Hydrol. **68**(3–4), 239–268 (2004)
Celia, M.A., Bouloutras, E.T., Zarba, R.L.: A general mass-conservative numerical solution for the unsaturated flow equation. Water Resour. Res. **26**(7), 1483–1496 (1990)
Claramunt, K., Cònsul, R., Pérez-Segarra, C.D., Oliva, A.: Multidimensional mathematical modeling and numerical investigation of co-flow partially premixed methane/air laminar flames. Combust. Flame **137**(4), 444–457 (2004)
Guarracino, L., Quintana, F.: A third-order accurate time scheme for variably saturated. Commun. Numer. Meth. Eng. **20**(5), 379–389 (2004)
Hairer, E., Nörsett, S.P., Wanner, G.: Solving ordinary equations I, Nonstiff problems, 2nd edn, Springer-Verlag, Berlin, 528 pp (2000)
Kavetsky, D., Binning, P., Sloan, S.W.: Adaptative time stepping and error control in a mass conservative numerical solution of the mixed form of Richards equation. Adv. Water Resour. **24**(6), 595–605 (2001)
Lehmann, F., Ackerer, P.: Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media. Transp. Porous Media **31**(3), 275–292 (1998)
Liao, W., Zhu, J., Khaliq, A.Q.M.: An efficient high-order algorithm for solving systems of reaction-diffusion equations. Numer. Methods Part. Differ. Equ. **18**(3), 340–354 (2002)
Natividad, M.C., Stynes, M.: Richardson extrapolation for a convection-diffusion problem using a Shishkin mesh. Appl. Numer. Math. **45**(2-3), 315–329 (2003)
Richards, S.A.: Completed Richardson extrapolation in space and time. Commun. Numer. Methods Eng. **13**(7), 573–582 (1997)
Richardson, L.F.: The approximate arithmetical solution by finite difference of physical problems involving differential equations, with an application to the stress in a masonry dam. Philos. Trans. Roy. Soc. London **210**(A), 307–357 (1910)
Richardson, L.F.: The deferred approach to the limit. I: single lattice. Philos. Trans. Roy. Soc. London **226**(A), 299–349 (1927)
Rubin, J.: Transport of reacting solutes in porous media: relation between mathematical nature of problem formulation and chemical nature of reaction. Water Resour. Res. **19**(5), 1231–1252 (1983)
Schneid, E., Knabner, P., Radu, F.: A priori error estimates for a mixed finite element discretization of the Richards' equation. Numer. Math. **98**(2), 353–370 (2004)
Shampine, L.F.: Local error estimation by doubling. Computing **34**(2), 179–190 (1985)
Steefel, C.I., McQuarrie, K.T.B.: Approaches to modelling of reactive transport in porous media. In: Lichtner, P.C., Steefel, C.I., Oelkers, E.H. (eds.) Reactive Transport in Porous Media. Reviews in Mineralogy, vol. 34, Mineralogical Society of America, Washington, pp. 82–129 (1996)
Sun, S., Wheeler, M.F.: Discontinuous Galerkin methods for coupled flow and reactive transport problems. Appl. Numer. Math. **52**(2–3), 273–298 (2005)
Valocchi, A.J., Street, R.L., Roberts, P.V.: Transport of ion-exchanging solutes in groundwater: chromatographic theory and field simulation. Water Resour. Res. **17**(5), 1517–1527 (1981)
van Genuchten, M.Th.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. **44**(5), 892–898 (1980)
Yeh, G.T., Tripathi, V.S.: A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components. Water Resour. Res. **25**(1), 93–108 (1989)
Zienkiewicz, O.C., Zhu, J.Z.: Superconvergent patch recovery and a posteriori error estimates. Part 2: error estimates and adaptivity. Int. J. Numer. Methods Eng. **33**(7), 1365–1382 (1992)

Annexe 7. Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems – the MoMaS benchmark case

Comput Geosci (2010) 14:483–502
DOI 10.1007/s10596-010-9178-2

ORIGINAL PAPER

Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems—the MoMaS benchmark case

Jérôme Carrayrou · Joachim Hoffmann · Peter Knabner · Serge Kräutle · Caroline de Dieuleveult · Jocelyne Erhel · Jan Van der Lee · V. Lagneau · K. Ulrich Mayer · Kerry T. B. MacQuarrie

Received: 15 September 2009 / Accepted: 3 February 2010 / Published online: 2 March 2010

Abstract Although multicomponent reactive transport modeling is gaining wider application in various geoscience fields, it continues to present significant mathematical and computational challenges. There is a need to solve and compare the solutions to complex benchmark problems, using a variety of codes, because such intercomparisons can reveal promising numerical solution approaches and increase confidence in the application of reactive transport codes. In this contribution, the results and performance of five current reactive transport codes are compared for the 1D and 2D subproblems of the so-called easy test case of the MoMaS benchmark (Carrayrou et al., Comput Geosci, 2009, this issue). This benchmark presents a simple fictitious reactive transport problem that highlights the main numerical difficulties encountered in real reactive transport problems. As a group, the codes include iterative and noniterative operator splitting and global implicit solution approaches. The 1D easy advective and 1D easy diffusive scenarios were solved using all codes, and, in general, there was a good agreement, with solution discrepancies limited to regions with rapid concentration changes. Computational demands were typically consistent with what was expected for the various solution approaches. The differences between solutions given by the three codes solving the 2D problem are more important. The very high computing effort required by the 2D problem illustrates the importance of parallel computations. The most important outcome of the benchmark exercise is that all codes are able to generate comparable results for problems of significant complexity and computational difficulty.

Keywords MoMaS · Benchmark · Code intercomparison · Numerical methods for reactive transport · Direct substitution approach (DSA) · Differential and algebraic equations (DAE) · Sequential iterative approach (SIA) · Sequential noniterative approach (SNIA)

J. Carrayrou (✉)
Institut de Mécanique des Fluides et des Solides,
Laboratoire d'Hydrogéologie et de Géochimie
de Strasbourg, University of Strasbourg,
UMR 7517 UdS-CNRS, Strasbourg, France
e-mail: carrayro@imfs.u-strasbg.fr

J. Hoffmann · P. Knabner · S. Kräutle
Department of Mathematics, University
of Erlangen-Nuremberg, Erlangen, Germany

C. de Dieuleveult · J. Erhel
INRIA Rennes, Campus de Beaulieu, 35042 Rennes, France

C. de Dieuleveult
ANDRA, Parc de la Croix-Blanche,
92298 Châtenay-Malabry, France

J. Van der Lee · V. Lagneau
Mines ParisTech, 35 rue Saint Honoré,
77305 Fontainebleau Cedex, France

K. U. Mayer
Department of Earth and Ocean Sciences, University
of British Columbia, Vancouver, BC, Canada
e-mail: umayer@eos.ubc.ca

K. T. B. MacQuarrie
Department of Civil Engineering, University
of New Brunswick, Fredericton, NB, Canada, E3B 6E8
e-mail: ktm@unb.ca

1 Introduction

Modeling reactive transport in porous media requires the solution of a coupled set of equations describing the transport of mobile chemical species together with a variety of geochemical reactions [43]. Since initiation of research in this field, reactive transport modeling has been recognized as a problem that may lead to significant mathematical and numerical difficulties. These difficulties originate from numerous challenges related to the solution of each operator (i.e., transport and chemistry) and the coupling of the operators used to evaluate the transport and reaction phenomena. As a result, a body of literature is developing that is devoted to the verification and validation of reactive transport models. In addition, several authors have conducted studies focusing on the performance assessment of reactive transport models and related solution methods. One can distinguish between four cases for these studies:

- Method evaluation based on theoretical considerations
- Comparisons of numerical results with exact or quasiexact solutions
- Intercomparisons of results obtained from two or more numerical methods
- Validation of numerical models based on comparing simulation results with experimental data

A key paper based on theoretical comparisons of solution approaches was presented by Yeh and Tripathi [54]. In this paper, the methods for coupling transport and chemistry were studied, and sequential and global methods were compared with respect to memory requirements and computing time, and calculations were performed based on estimates of the number of unknowns and the number of operations associated with each method. The literature devoted to the evaluation of errors on transport–chemistry (T–C) coupling follows a similar approach. In several contributions (e.g., Valocchi and Malmstead [47], Kaluarachchi and Morshed [20], Barry et al. [2, 3], Leeming et al. [24], Kanney et al. [21], Carrayrou et al. [9]), a variety of methods were evaluated by comparing mass balances obtained using the sequential approaches with exact mass balances.

Numerous verification studies have been performed by comparing numerical and exact analytical solutions. Unfortunately, the problems handled by analytical solutions are highly simplified and do not allow a full evaluation of the capabilities of multicomponent reactive transport codes. Available analytical solutions are typically restricted to 1D transport of a single species in homogeneous media (e.g., Van Genuchten and Wierenga [49], Selim and Mansell [38], Van Genuchten [48], Carnahan and Remer [5]). Some studies deal with 2D and 3D transport [45], and a few attempts have been made to include more complex chemical reaction networks. For example, Toride et al. [46] considered a two-site sorption model present in both mobile and immobile domains. However, analytical solutions are generally limited to homogeneous and unidirectional flow fields, and the geochemical system involves only one or two reactions described either by isotherms or by first-order rate expressions. In reality, flow systems are not restricted to one spatial dimension but may require 2D [14] or 3D [18] spatial discretizations, often further complicated by physical and chemical heterogeneities [4] or fractures [30]. The chemical reaction network may include instantaneous equilibrium reactions [53], kinetic processes [37], or a mixed reaction network (e.g., Mayer et al. [27]), subject to a high degree of coupling and nonlinearity. Processes may include mineral weathering and formation [25], biological phenomena [32], radioactive decay [15], competitive sorption and ion exchange [44], and isotope fractionation [33] and may involve more than 200 chemical species (e.g., Bain et al. [1]).

Model validation can be attempted by comparing numerical results with experimental data. For example, van Genuchten et al. [50] evaluated a reactive transport model based on experimental data that describes transport and nonlinear sorption of trichlorophenoxyacetic acid. Validation of reactive transport models is an important task; however, the a priori verification of the numerical code is still required because it needs to be demonstrated that the numerical code solves the governing equations correctly and accurately. Comparisons of simulation results to experimental data alone do not provide a suitable tool for model verification. This approach does not allow distinguishing between differences that are due to an incorrect implementation of the governing equations, discrepancies associated with an incomplete or faulty conceptual model, or deviations associated with experimental and analytical uncertainties.

Based on these limitations, a suitable avenue for model verification appears to be the intercomparison of numerical results. This intercomparison involves the independent solution of the same problem using a variety of models and/or numerical techniques. One of the main advantages of this method is that complex systems that are more representative of real world reactive transport problems can be considered. The intercomparison of numerical results also has some disadvantages, specifically that the *true* solution of the problem

is not known; however, obtaining the same or very similar results with a variety of computer codes, which are based on different methods and implementations, provides increased confidence in the accuracy of the codes and the field of reactive transport modeling in general.

Despite these obvious benefits, very few model intercomparisons have been published to date. Freedman and Ibaraki [17] compared different solution approaches to model redox processes by comparing the two codes DYNAMIX and DART. De Windt et al. [14] present an intercomparison of the reactive transport codes CASTEM, CHEMTRAP, PHREEQC, and HYTEC for the simulation of oxidation, dissolution, and transport of uranium. The intercomparison presented by De Windt et al. [14] involves a relatively complex chemistry geochemical system and a 2D flow field. In addition, there are very few comparisons that provide information about the performance of the numerical methods used. The literature devoted to the comparison of sequential and global approaches for T–C coupling [13, 16, 35, 36, 39, 42, 43] provides some discussion that is mostly qualitative in nature. Reeves and Kirkner [34] provide the computing times required for the solution of a 1D problem with sorption of one, two, or three components for a number of methods. In these studies, comparisons are typically based on the same mesh size and/or the same time step, despite the fact that each method requires its own time step and mesh size.

Hence, the literature devoted to comparison of numerical solutions for reactive transport models is subject to some limitations, such as

- low degree of complexity
- lack of performance evaluation
- low number (two or three codes) of simultaneous comparisons

The reactive transport benchmark of MoMaS has been designed to help fill these gaps. The benchmark provides a high degree of complexity and nonlinear coupling and provides a platform that allows focusing on the comparison of methods and implementations by ensuring that all participants use the same model. The reaction network is synthetic in nature, removing the dependence on the formulation of activity corrections or database dependencies. Results are thus strictly identical from a chemical perspective. The objectives of this benchmark are then to compare the numerical methods and their implementations.

The first objective is to analyze the ability of the different methods to solve the various benchmark tests. We investigate three classes of numerical coupling: sequential noniterative approach (SNIA) based on transport operator splitting and no iteration between transport and chemistry, sequential iterative approach (SIA) based on an implicit scheme and fixed-point iterations for nonlinear coupling of transport and chemistry, and global methods based on an implicit scheme and Newton iterations for nonlinear coupling. We do not investigate SNIA methods based on an explicit scheme.

The second objective is to provide a measure for computational efficiency. Twenty years ago, Yeh and Tripathi [54] concluded that "Those models that use the DAE approach or the DSA require excessive CPU memory and CPU time. They can only remain as a research tool for one-dimensional problems." We design challenging 1D and 2D test cases in order to check if modern, global approaches can compete with sequential approaches. We compare three implementations of the global approach, which differ by the number of coupled unknowns, in order to measure the impact of a reduction of unknowns. The efficiency is strongly related to the numerical coupling but also to the discretization schemes, to the solution algorithms, and to the implementation. For example, various strategies have been implemented to control the time step and to control the convergence of nonlinear iterations. We do not aim at ranking the methods and the codes. Indeed, the conclusions are valid only for the test cases used, some of the codes are still under development, and the computers used are not the same. Despite these limitations, we attempt to draw conclusions regarding performance of the methods with general relevance.

The third objective is to provide a measure for the accuracy of the numerical results. The comparison must be global but must also highlight some local key features such as a peak of concentration. Accuracy can be analyzed qualitatively by using, for example, visualization tools. In order to derive a quantitative measure, it is necessary to define a reference solution. Again, we try to draw some general conclusions, based on the results of the test cases.

This paper presents results from five different research teams using five different approaches: SNIA with operator splitting, SIA, and three variants of global approaches. This contribution presents a synthesis of the results obtained by the five codes. We use four test cases, from the so-called Easy test case collection of the MoMaS reactive transport benchmark. Additional simulation results for these test cases and other test cases [10] are documented in the contributions by the individual participants [6, 12, 19, 23, 26].

We first describe the reactive transport model used for designing the benchmark. Then, we briefly present the five codes used, along with a short description of their main features. Before presenting the results,

Table 1 Equilibrium table for the easy test case

	X_1	X_2	X_3	X_4	S	K
C_1	0	−1	0	0	0	1.00E−12
C_2	0	1	1	0	0	1
C_3	0	−1	0	1	0	1
C_4	0	−4	1	3	0	0.1
C_5	0	4	3	1	0	1.00E+35
CS_1	0	3	1	0	1	1.00E+6
CS_2	0	−3	0	1	2	1.00E−01
Total (m L^{-3})	T_1	T_2	T_3	T_4	TS	
Initial for medium A	0	−2	0	2	1	
Initial for medium B	0	−2	0	2	10	
Injection $t \in [0; 5{,}000]$	Imposed total concentration at inflow boundary					
Inflow for 1D	0.3	0.3	0.3	0		
Zone 1 for 2D	0.3	0.3	0.3	0		
Zone 2 for 2D	0.3	0.3	0.3	0		
Leaching $t \in [5{,}000; \ldots]$	Imposed total concentration at inflow boundary					
Inflow for 1D	0	−2	0	2		
Zone 1 for 2D	0	−2	0	2		
Zone 2 for 2D	0	−2	0	2		

we describe the methodology used for achieving the objectives of comparison. Finally, we discuss the results and provide some concluding remarks.

2 Reactive transport model

Reactive transport is described using the advection–dispersion equation with reactions subject to the instantaneous equilibrium assumption:

$$\omega \frac{\partial \left(T_{\mathrm{M}_j} + T_{\mathrm{F}_j}\right)}{\partial t} = -\nabla \left(\omega u T_{\mathrm{M}_j}\right) + \nabla \left(\overline{\overline{D}} \times \nabla T_{\mathrm{M}_j}\right) \quad (1)$$

Where t is the time, u is the pore water velocity, T_{M_j} is the total mobile concentration for each component and T_{F_j} is the total immobile concentration. $\overline{\overline{D}}$ is the dispersion tensor, and ω is the porosity. Chemical reactions give the relations between T_{M_j} and T_{F_j} by the way of mass action laws and conservation equations.

The chemical phenomena are summarized in form of an equilibrium tableau in Table 1. The reactions involve four aqueous components and one immobile component, leading to the formation of five aqueous and two adsorbed secondary species. A characteristic of this chemical system is that it contains very high stoichiometric coefficients: from −4 to 4 for component X_2; and equilibrium constants encompassing an extreme range from 10^{-12} for C_1 to 10^{35} for C_5.

One-dimensional and 2D domains were studied. For both cases, the domains are heterogeneous both in terms of hydrodynamic and chemical properties (see Fig. 1). The domains are composed of two media: Medium A is highly permeable, with low porosity and low reactivity, whereas medium B has a low permeability with high porosity and high reactivity. A complete

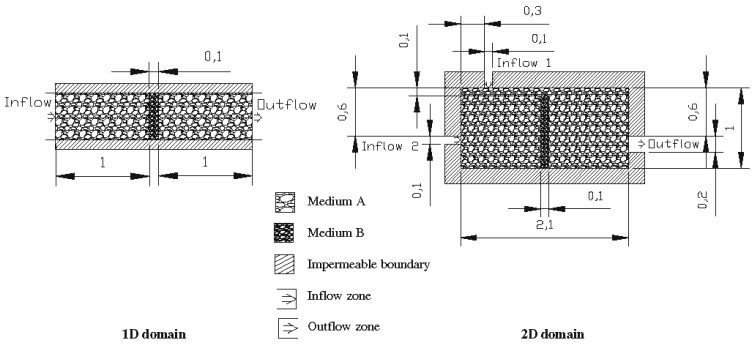


Fig. 1 Scheme of the 1D and 2D domains

description of the exercise can be found in Carrayrou et al. [10].

3 Numerical methods and codes

Brief summaries of the key features of the codes used by the benchmark participants are presented below with a focus on the most significant differences between implementations. Table 2 provides an overview of the key characteristics of the codes: The first row entries describe the method of coupling between transport and chemistry operators; the second row entries introduce the formulation for advection and dispersion operators; the third row entries describe the method used for spatial discretization; the fourth row entries represent the time discretization used; in the fifth row, the method used to linearize the chemical system is provided; the sixth row entries describe the convergence criteria used for linearization (all criteria have been tested and chosen sufficiently small to have no influence on the accuracy of the proposed solutions); and the last row represents the method used for the solution of the linearized system of equations. For a more detailed description of the codes, we refer to the individual articles in this special issue. Although this work is devoted to a comparison of numerical methods implemented in the participating reactive transport codes, the general capabilities of the codes are presented for completeness and to provide additional perspective (Table 3).

3.1 GDAE1D

This code is based on a method of lines in combination with a global approach in order to solve the partial differential algebraic equations (DAE) involving transport and chemistry [12, 13]. In the current version, spatial discretization is achieved by a classical finite volume method, with upwinding for advection and centered spatial discretization for dispersion. The design of the mesh uses constant spatial discretization intervals. The resulting DAE are solved by an external, robust, and efficient DAE solver. Time discretization is performed by a multistep implicit scheme: a backward differentiation formula (BDF) with variable order and variable time step. BDF is used in connection with a modified Newton method in order to deal with nonlinearity. The sparse linear systems are solved by a direct method, a multifrontal Gaussian elimination with pivoting. Symbolic factorization and renumbering for fill-in reduction are performed once by using the matrix structure. Due to the connection between BDF and Newton's method, the Jacobian matrix is updated only when necessary and the time step is controlled to ensure both convergence of Newton's method and the accuracy of the scheme. The main computational cost is associated with the factorization of the Jacobian matrix and the solution of the triangular system of equations. For large computational domains, it is necessary to decrease the computational cost. Several issues will be addressed in future versions: the spatial grid will be nonuniform; the tolerance thresholds in the DAE solver will be tuned; and the substitution approach will be applied in the linear system in order to reduce the number of unknowns. For the benchmark exercise, 600 cells were used for the 1D advective case, while 400 cells were used for the 1D dispersive case. Small tolerance thresholds were specified to the DAE solver.

3.2 Code of Hoffmann et al.

This solution method reduces the size of the nonlinear system and, thus, the required computational resources. The system of equations, consisting of partial (PDEs) and ordinary differential equations (ODEs) for the mobile and immobile species and nonlinear advection equations (AEs) describing local equilibria, is transformed by (a) taking linear combinations between the differential equations, (b) the introduction of a new set of variables, i.e., a linear variable transformation, and (c) the elimination of some of the new variables by substituting local equations, such as AEs and ODEs, into the PDEs. Application of (a) and (b) leads to a decoupling of the linear PDEs; this decoupling in combination with (c) leads to a reduction of the size of the nonlinear system (see Kräutle and Knaber [22], Hoffmann et al. [19], and the references therein for details). The system of equations is handled in the spirit of a global implicit approach (one-step method) and avoids operator splitting. However, the substitution of the local equations does not, as is the case for other direct substitution approaches, destroy the linearity of the transport term. The algorithm was implemented using a software kernel for parallel computations involving PDEs, called "M++." M++ itself is an object oriented code based on C++. The code is implemented for 2D problems and uses finite elements on unstructured grids. The nonlinear system of equations is linearized using Newton's method and solved using a preconditioned BiCGStab algorithm. For the solution of the flow problem, mixed hybrid finite elements are used. For the flow computation in the 2D case of this benchmark, Brezzi-Douglas-Marini elements of order one were used. This method guarantees an accurate solution of the flow problem despite the significant permeability contrast between the two media. To facilitate

Table 2 Summary of the main features of the compared codes

Method	SPECY	HYTEC	MIN3P	Hoffmann et al.	GDAE1D
Transport–chemistry coupling	SNIA no iteration	SIA fixed-point iterations	DSA	Reduced-global ODE approach	DAE
Advection–dispersion OS	Yes	No	No	No	No
Spatial discretization	Discontinuous finite element for advection and mixed finite element for dispersion	Finite volume and Voronoï mesh	Finite volume, upwinding, centered, or flux limiter for advective term	Finite element and finite volume upwinding	Finite volume upwinding
Time discretization	Explicit for advection implicit for dispersion and constant time step	Implicit adaptive time step	Implicit adaptive time stepping	Implicit adaptive time step	Implicit adaptive order adaptive time step
Linearization	Chemistry only: Positive continuous fraction as preconditioner and Newton–Raphson with restricted search domain	Chemistry only: Newton–Raphson with relaxation factor	Modified Newton substitution of variables	Newton reduction of variables and substitution	Newton no substitution
Convergence criteria	For chemistry only: Relative error on mass balance of 10^{-12}	Relative criteria on fixed concentration of 10^{-8}	Relative criteria on concentration update: relative error of $\Delta \ln c < 10^{-8}$	Relative and absolute criteria on the residual of Newton's method. Relative error of 10^{-6}; Residual of 10^{-10}	
Resolution of linear systems	Direct multifrontal solver	Preconditioned GMRES	Preconditioned BiCGStab algorithm	Preconditioned BiCGStab algorithm	Direct method, a multifrontal Gaussian elimination with pivoting

Table 3 Summary of the main phenomena and models included into the compared codes

Phenomena	SPECY	HYTEC	MIN3P	Hoffmann et al.	GDAE1D
Flow field	No, calculated separately	Constant or transient, unsaturated (Richards equation)	Yes	Constant or calculated by solving Richards equation	No
Unsaturated flow	No	Yes	Yes	Yes	No
Multiphase flow	No	No	No	Implementation in progress	No
Gas phase transport	No	No	Yes	No	No
Variation of porosity	No	Yes	Yes	No	No
Aqueous-gas exchange	Henry's law, included into mass action laws	Henry's law, included into mass action law	Yes	Phase diagrams given following molar fractions	No
Activity correction	Debye–Hückel Güntelberg Davies	Debye–Hückel, extended Debye–Hückel, Davis, B-dot, SIT	Extended Debye–Hückel Davis Pitzer	No	No
Precipitation dissolution	Yes Instantaneous equilibrium and kinetic	Yes Equilibrium and kinetic	Yes Kinetic formulation	Yes, instantaneous equilibrium and kinetic, using formulation with complementarity condition	No
Ion exchange	Yes	Yes	Yes	Yes	Yes
Surface complexation	Constant capacity model Diffuse layer model Stern model Triple layer model	Constant capacity, diffuse layer, and triple layer models	Nonelectrostatic adsorption model	Sorption according to law of mass action with surface complex No electrostatic correction	Nonelectrostatic adsorption model
Kinetic chemistry	Yes Model to be defined by user	Yes: distance from equilibrium, Monod, energy term, catalysis/inhibition (all reactions: aqueous, surface or mineral)	Yes (intra-aqueous kinetic reactions and dissolution precipitation)	Yes law of mass action	No
Biologic	Not specific but included into the kinetic module	Specified through kinetic reactions	Can be specified through kinetic reactions	Monod model	No

fair comparison with the other models, the code was run on a single processor.

3.3 SPECY

SPECY uses a noniterative operator splitting scheme for T–C coupling and for advection and dispersion [8]. Each operator is solved independently using specifically tailored methods: advection is solved using discontinuous finite elements [40]; dispersion is tackled with mixed hybrid finite elements; and equilibrium chemistry is solved using a combined algorithm based on the Newton–Raphson technique and the positive continuous fraction method [7]. The key feature of this code is the use of specific methods to solve each part of the reactive transport equation. Solving the advective part using discontinuous finite elements provides an excellent description of very sharp fronts and eliminates numerical diffusion and nonphysical oscillations. Solving the dispersion term with mixed hybrid finite elements provides an exact mass balance for each element of the mesh and allows the use of a nondiagonal dispersion tensor. The algorithm developed for solving the equilibrium chemistry ensures the convergence of the method for all cases and provides fast convergence for most cases. To optimize computational performance, we used the largest time step allowed by SPECY. This constant time step length is determined by a Courant–Friedrich–Levy stability criterion equal to one. The reader is refereed to Carrayrou [6] for additional details on the code formulation and its application to the MoMaS reactive transport benchmark.

3.4 HYTEC

HYTEC is a reactive transport model that integrates a wide variety of features and options that have evolved, after more than a decade of development, to a widely used and versatile simulation tool [51]. Solution capabilities for biogeochemistry are provided by the code CHESS (http://chess.ensmp.fr). The model accounts for many commonly encountered processes including interface reactions (surface complexation with electrostatic correction and cation exchange), precipitation and dissolution of solid phases (minerals and colloids), organic complexation, redox and microbial reactions, etc. All reactions can be modeled using a full equilibrium, a full kinetic, or a mixed equilibrium–kinetic approach. Thermodynamic data are taken from the database developed by the Common Thermodynamic Database Project.

The hydrodynamic module of HYTEC is adapted for hydrodynamic conditions commonly encountered in the laboratory or in the field. The code allows for unsaturated media, variable boundary conditions, sinks, and sources [52]. HYTEC searches for an accurate solution to the multicomponent transport problem using an iterative, sequential, so-called strong coupling scheme. Strong coupling permits variable hydrodynamic parameters as a function of the local chemistry. For example, the porosity of a porous medium reduces after massive precipitation of newly formed mineral phases, which modifies the water flow paths and transport parameters, e.g., diffusion coefficients: HYTEC solves this interdependency accurately, which makes the tool particularly useful for, e.g., cement alteration at long timescales (e.g., storage of wastes and performance assessment).

Application domains of HYTEC are numerous and include soil pollution, acid mine drainage, in situ leaching of copper or uranium, radioactive waste disposal (performance assessment, near- and far-field processes), and storage of greenhouse gases. Other applications concern the evolution and degradation of (geo)materials such as ashes, concrete, and cements; the latter often being simulated by a typical CEM-I cement, but more sophisticated models for cements can be used including sorption on primary or secondary calcium silicate hydrate phases, carbonation, and sulfatation of the material. The strong coupling approach as outlined above makes HYTEC particularly useful for the modeling of long-term leaching of solidified wastes.

Efforts to develop, test, and validate the HYTEC model largely exceed the scope of a single laboratory and the timescale of a Ph.D. thesis. The Reactive Transport Consortium (*Pôle Géochimie-Transport* [PGT], http://pgt.ensmp.fr) is a national research project with the objective of creating a long-term framework for the development of reactive transport models, reference studies, and new application domains. Already operational for several years, the collaborative efforts within the PGT allowed to make considerable progress in the domain of reactive transport modeling.

3.5 MIN3P

MIN3P is designed to simulate general flow and reactive transport problems in variably saturated media for 1D to 3D systems. The flow solution is based on Richard's equation, and transport of solute is simulated using the advection–dispersion equation [28]. Gas transport is by diffusion only in the standard version of the code [28] or by advection and diffusion within the framework of the dusty gas model [31]. Geochemical processes included are aqueous complexation, mineral dissolution–precipitation, intra-aqueous kinetic

reactions, gas dissolution, ion exchange, surface complexation, and linear sorption. All reactions considered in the simulations can be specified through a database. The code has been used for a wide range of applications in the field of contaminant transport (e.g., Mayer et al. [27]) and groundwater remediation (e.g., Mayer et al. [29]). The code was also used for investigation of redox stability in crystalline rock formations that may be considered for deep geologic repositories for nuclear waste [41].

The solution of the governing equations is based on the global implicit method, in which the reaction equations are directly substituted into the transport equations, known as the "direct substitution approach" (DSA) [54]. Spatial discretization is performed using a control volume method with half-cells on the boundary. The code uses implicit time weighting and provides a choice of various spatial weighting schemes for advective transport, including upstream weighting, which was used for the current simulations. The governing equations are linearized using a modified Newton's method with variable time stepping; a sparse iterative solver is used for the solution of the linearized matrix equations (see Mayer and MacQuarrie [26], for additional details). For the easy test case presented here, the code was used without any modifications.

4 Methodology of comparison

In order to interest as many research teams as possible and to extend the applicability of the benchmark to a wide variety of methods, the hydrodynamic flow system has been kept straightforward, with only two media and a simple 1D or 2D geometry. For the same reason, the chemical system has been simplified in the sense that activity corrections have been neglected and that sorption reactions do not include electrostatic correction terms. On the other hand, the benchmark has been designed to ensure a high degree of numerical difficulty: physical and chemical heterogeneities are significant, chemical phenomena are strongly coupled and nonlinear, and concentration gradients induced by external forcing due to changes in boundary conditions are substantial.

In this contribution, we focus on a comparison of the results for the easy test case, both for 1D and 2D computational domains, and for the advective and dispersive scenarios. All the five codes have results for the 1D test cases; on the other hand, only three codes give results for the 2D advective test case and only two codes for the 2D dispersive test case; similar results for the 2D test cases can also be found in de Dieuleveult's Ph.D. thesis [11].

We first measure the computational complexity of the codes; since most of them use an adaptive time step, we only measure the CPU time as a function of the number of cells. The CPU time is specified in terms of a system-independent CPU unit, which is defined in the paper introducing the benchmark exercise [10]. Although the CPU time comparison is intended to provide an objective performance-based measure of model and method applicabilities for the various test cases, this method has some limitations. Some codes are in the process of development (GDAE1D de Dieuleveult and Erhel [12], Hoffmann et al. [19]) and only include a limited chemical reaction network, whereas other programs (SPECY, HYTEC, and MIN3P) can handle general and complex reaction networks; in these codes, chemistry can be specified from a database, not only greatly increasing model flexibility but also generating computational overhead (see Table 3). In addition, providing a measure of the computational effort independent of computing hardware and compiler software is quite difficult. The computational complexity must therefore be considered qualitative. For further information on the variability of CPU times as a function of system parameters, we refer to the contribution of de Dieuleveult and Erhel [12].

In the following, the accuracy of the codes is compared. Since the methods used are different, they require different spatial and temporal discretizations to obtain a solution of the same accuracy. Therefore, CPU as a function of grid size should not be assessed in isolation. We could compare the accuracy of codes by using the same number of cells in all of them. We choose a different strategy and compare the accuracy of codes by using the same normalized CPU time for all of them. Maximum allowed computing times are specified for each test case investigated. For the easy test case presented here, the following maximum CPU units were imposed: 3,500 units for 1D advective case, 2,000 units for 1D dispersive case, 10,000 units for 2D advective case, and 10,000 units for 2D dispersive case. Again, this exercise has some limits, but it provides some useful information.

Since the benchmark is designed for handling complex models, there is no analytical reference solution. Since the test cases are synthetic, there is no experimental reference solution. Therefore, it is difficult to derive a quantitative comparison. For the 1D test cases, reference solutions are calculated using fine grids and small time steps, providing a basis for accuracy measurement. An example of this approach is given by Carrayrou [6]. The validity of these reference solutions has been

controlled by successive mesh and time step refinements and by comparison with refined solution from the other codes. Then, we use the reference solution to define an error criteria based on a L2 norm. The norm (L2) is calculated for the studied species ($C_{\text{calculated}}$) over the interval (noted "L"), which can be either the space domain (x varying from 0.0 to 2.1 in 1D case, x varying from 0.0 to 2.1, and y varying from 0.0 to 1.0 in 2D case) or the simulation time (time from 0.0 to 6,000.0). A relative error or deviation between the solutions can be quantified by the L2 norm, which is defined by Eq. 2:

$$L2 = \sqrt{L \times \sum_{\substack{\text{All the } L \\ \text{discretisation}}} \left[\frac{\Delta L_i}{L} \left(C_{i,\text{calculated}} - \sum_{\substack{\Delta L_i \text{ discretisation} \\ \text{on reference}}} \frac{dL_{j,\text{ref}} \times C_{j,\text{reference}}}{\Delta L_i} \right) \right]^2} \tag{2}$$

In Eq. 2, ΔL is the discretization used by the calculated solution and $dL_{j,\text{ref}}$ is the discretization used by the reference solution over ΔL_i.

For the 2D test cases, it was not possible to define a reliable reference solution because computational requirements were too high for a very refined mesh. In order to compute an L2 norm, we used the most refined computation as reference.

This criterion gives a global quantitative comparison of accuracy. However, since there are many species, with concentrations varying in space and time, it is difficult to represent and to analyze all the results. The global quantitative comparison gives some information but does not highlight some local key points. In order to compare the local accuracy of the codes, we select representative results that focus on key difficulties of the benchmark and, at the same time, highlight the most significant differences between the five codes. Thus, we compare the results given by the codes for some specific species at some specific time or location. The purpose of this comparison is to analyze if a code can compute an accurate solution for a specific pollutant or near a pumping well.

5 Results

5.1 Computational complexity

To illustrate the computational complexity of the various codes, we plot the normalized CPU times as a function of the number of cells in the mesh. Results for the 1D advective and dispersive test cases are presented in Figs. 2 and 3, respectively. Results for the 2D advective test case are presented in Fig. 4.

As expected, the computational complexity of all codes is characterized by a linear log–log relationship between CPU time and mesh size, independent of the test case considered. It appears that all codes have the same slope for the 1D test cases (except HYTEC for the 1D advective test case). For the 1D advective and dispersive test cases (Figs. 2 and 3), well-known results are confirmed: the SNIA (SPECY) is faster than other methods, for a fixed number of cells. However, as suggested by Saaltink et al. [35], implementations of the DSA approach (e.g., MIN3P, Hoffmann et al.) can lead to competitive CPU performance. The new reduction scheme developed by Kräutle and Knabner [22] (see also Hoffmann et al. [19]) decreases further the computational complexity. Despite the use of a global approach, this implementation shows equivalent or lower CPU times than required by all other codes. Moreover, it must be underlined that this code uses a 2D discretization to emulate a 1D domain. This method is more CPU time consuming than solving a 1D problem. Global methods appear very competitive for the 2D advective test case. Extrapolating the performance data for each of the three codes in Fig. 4 shows that for a mesh with the same number of cells, the CPU requirements for the code by Hoffmann et al. are more

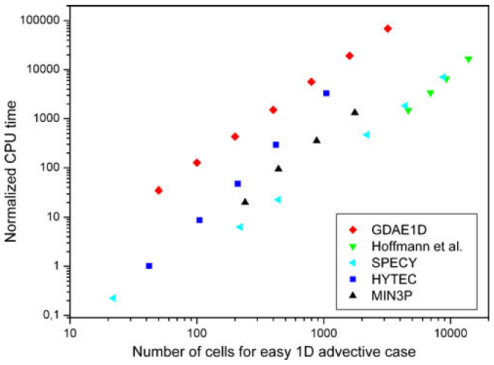


Fig. 2 Normalized computing times as a function of discretization for the 1D easy advective test case

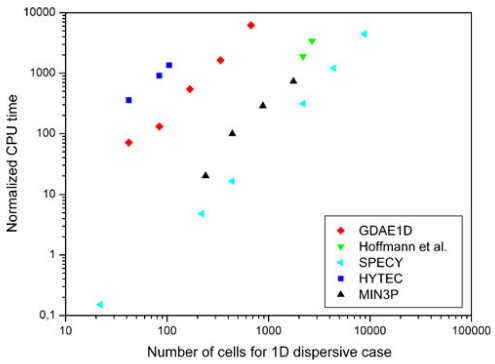


Fig. 3 Normalized computing times as a function of discretization for the 1D easy dispersive test case

than five times lower than the CPU times of the two other codes.

However, we emphasize that this measure does not provide insight for accuracy. So, now we present a comparison of accuracy, with all the codes using approximately the same normalized CPU time.

5.2 Accuracy for 1D easy advective test case

The requirement to limit CPU times to no more than 3,500 CPU units led to a range of spatial discretizations for the various codes. GDAE1D used 600 uniform cells, while HYTEC was run with 1,073 uniform cells. The SPECY and MIN3P simulations were conducted with nonuniform grids. The discretization in the low-permeability zone in the center of the domain (medium B) was refined by a factor of 2; SPECY and MIN3P employed 6,400 and 1,760 cells, respectively. Hoffmann et al. used a 2D discretization to emulate the 1D problem by replacing the 1D computational domain with a narrow 2D domain. A preadapted triangular mesh was used with different grid sizes in the two media: grid size h_1 in medium A and grid size h_2 in medium B with $h_1 = 4h_2$. The resulting mesh consists of 6,942 cells with 1,155 nodes in the x-direction. In medium A, the mesh has three nodes in the y-direction.

A global quantitative comparison between the results given by each code and the reference solution is performed using the L2 error norm (see Table 5). The reference solution is given by SPECY using a 8,200 cells mesh and a constant time step of 1.14×10^{-4}. All the codes provide similar error norms. The best results are obtained by GDAE1D, although the approach chosen by GDAE1D is computationally intensive and requires using a coarse grid to respect the specified CPU time criteria. The results provided by HYTEC leads to the second L2 norm. The results given by the code of Hoffmann et al. and by MIN3P lead to the third and fourth L2 norms.

This global criterion is not sufficient to compare accuracy. To compare local results for this test case, we have selected the concentration profile of the fixed component S at time 10. This profile is characterized by sharp concentration fronts with a very narrow peak located near the inlet of the domain (Fig. 5). This concentration peak is due to the disequilibrium created by the injection of species X_3. The influence of the more reactive medium B can be seen in the center of the domain, as indicated by the higher concentration of S. All codes produce very similar concentration profiles at

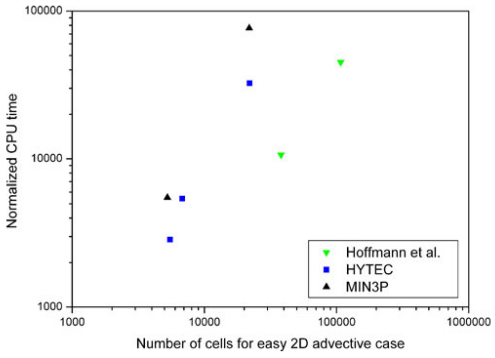


Fig. 4 Normalized computing times as a function of discretization for the 2D easy advective test case

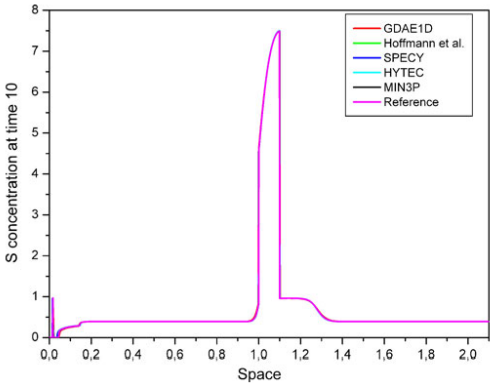


Fig. 5 Concentration profiles of solid component S at time 10 for the 1D easy advective test case

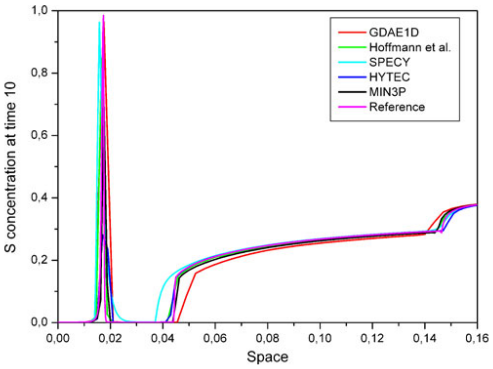


Fig. 6 Local concentration profiles of solid component S at time 10 for the 1D easy advective test case (subregion: $x = 0$ to $x = 0.16$)

the scale of the solution domain. More comprehensive results presented in the individual contributions for each code [6, 12, 19, 23, 26] confirm the good agreement for other chemical species.

However, Fig. 5 also reveals small discrepancies for the concentration peak near the domain inlet. Zooming into this region provides a sensitive measure for a more in-depth code comparison. The location and intensity of the peak at $x = 0.02$ (Fig. 6) provide a direct indication of coupling error or numerical diffusion. Figure 6 indicates that there are indeed small differences in the location of the concentration peak and the magnitude of the peak concentration.

Table 4 provides a quantitative assessment of these differences suggesting that all codes produce similar peak locations with a low standard deviation; however, the maximum concentrations calculated by the various codes are characterized by a wider range. Successive mesh and/or time step refinements performed using the various models indicate that for the exact solution of S, the peak concentration will exceed 0.9 (see Carrayrou [6]). The reference solution is a peak of 0.985.

Table 4 Location and peak amplitude for the first S peak at time 10 for the 1D easy advective test Case

	Location of the peak	S concentration of the peak
GDAE1D	0.0175	0.966
Hoffmann et al.	0.0167	0.852
SPECY	0.0158	0.968
HYTEC	0.0170	0.286
MIN3P	0.0175	0.725
Reference	0.0174	0.985
Mean	0.0169	0.759
Standard deviation	7.04×10^{-4}	0.283

Even if the intensity of the peak is low with HYTEC, its localization is good, and the rest of the curve fits well with the reference solution. Traditionally, one of the main advantages of operator splitting methods is that tailored numerical methods can be used for each operator, including exact transport schemes to minimize numerical diffusion [43]. This is confirmed by the results obtained using SPECY (Fig. 6, Table 4). However, this peak is shifted to the left. Moreover, the curve between $x = 0.04$ and $x = 0.15$ is far from the reference.

The closest peak location and intensity to the reference are computed by GDAE1D. Thus, this global method achieves high peak concentrations despite a relatively coarse discretization. This is probably due to a small error tolerance in the DAE solver, inducing small time steps. It seems to indicate that global methods can be implemented with a low degree of numerical diffusion. For GDAE1D, some differences can be seen on Fig. 6 between $x = 0.04$ and $x = 0.15$; they are probably due to a small number of grid cells.

5.3 Accuracy for 1D easy dispersive test case

For the 1D easy dispersive test case, the maximum normalized CPU time was set to 2,000 CPU units. To meet this criterion, GDAE1D used a uniform discretization with 400 cells, while the HYTEC simulation employed 137 uniform cells. As for the 1D advective case, the SPECY and MIN3P simulations used a nonuniform discretization with grid refinement in medium B (by a factor of 2). For the SPECY simulation, the domain is discretized into 5,800 cells, while the MIN3P simulation was based on a grid with 880 cells. Hoffmann et al. used a narrow 2D computational domain to describe the 1D system. However, unlike the 1D advective case, no grid refinement was performed, and a regular mesh with three nodes in the y-direction was specified. The resulting grid consists of 2,184 triangles with 547 nodes in the x-direction.

L2 error norms are given on Table 5. The reference solution is given by MIN3P using a 1,760 cells mesh and a time step limited to CFL = 1. Again, all codes provide similar norms. Code MIN3P leads to the smallest L2 norm, followed by GDAE1D, then the code Hoffmann et al., finally SPECY and HYTEC. Global approaches are efficient for dispersive problems, and the mesh used by MIN3P is the finest among other global codes.

For this case, local accuracy measurement is based on breakthrough curves for species C_2 at the outflow of the domain (Fig. 7). C_2 concentrations increase rapidly

Table 5 L2 norm for the different test cases calculated versus a reference solution

Case	SPECY	HYTEC	MIN3P	Hoffmann et al.	GDAE1D
1D advective Figure 5 Reference given by SPECY	7.67×10^{-2}	2.54×10^{-2}	5.40×10^{-2}	5.00×10^{-2}	1.75×10^{-2}
1D dispersive Figure 7 Reference given by MIN3P	2.63×10^{-2}	2.89×10^{-2}	1.25×10^{-3}	1.05×10^{-2}	6.92×10^{-3}

after approximately 300 time units, and they equal the composition of the injected solution, followed by a sharp drop due to the change of the inflow boundary condition (after 5,000 time units). The simulation results indicate that all codes consistently reproduce the increase and decrease of the C_2 concentration front (Fig. 7).

This dispersive test case provides a serious test for implementations based on the sequential approach. The short timescale of dispersive transport effectively leads to an increased solute flux with possible feedback on local chemistry from several neighboring cells. These types of problems are known to be prone to the introduction of coupling errors, while global methods are expected to perform well.

This hypothesis is confirmed by the results shown in Fig. 7, which indicate an excellent agreement between the different global approaches (GDAE1D, Hoffmann et al., and MIN3P). Discrepancies between these three codes are particularly small. On the other hand, the SIA and SNIA solutions show slight deviations. Minor differences are visible for the codes using the SIA and SNIA methods during the flushing period (>5,000 time units); however, it must be emphasized that the time frame displayed is less than 5 time units, while the total simulation period is 6,000 time units.

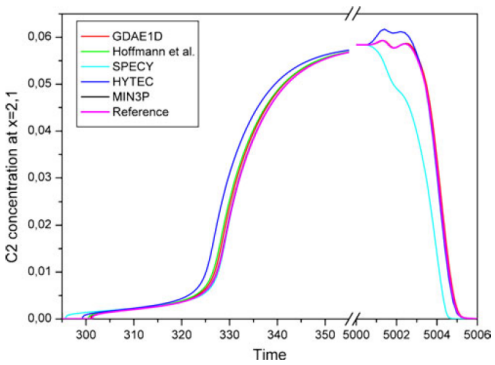


Fig. 7 Elution curve for species C_2 at $x = 2.1$ for the 1D easy diffusive test case

However, solutions obtained for refined grids (e.g., SPECY and Carrayrou [6]) converge toward the results obtained by the global methods, suggesting that errors are reduced by refining space and time.

5.4 Accuracy for the 2D easy advective test case

The 2D version of the easy advective test case was solved using three of the codes (HYTEC, MIN3P, and Hoffman et al.). Again, restricting the CPU time to a maximum of 10,000 units led to different spatial discretizations. Hoffmann et al. used a preadapted mesh with 38,016 triangles, refined in the fast velocity zone and near the outflow. The HYTEC solution used a grid with 8,840 cells (136×65) to comply with the CPU criterion. MIN3P employed a grid with 5,250 control volumes (105×50).

The concentration contours of component X_3 at time 1,000 offer a suitable means for comparison. Figure 8 clearly depicts high concentrations in the vicinity of the two injection zones, one located on the left boundary and the second located near the top of the model domain. High concentration regions are delineated by sharp fronts controlled by sorption and complexation reactions. In addition, the concentration distributions are significantly affected by the presence of medium B, which induces a deviation of the flow lines and a low concentration zone near the bottom of the domain.

Comparing the results demonstrates that all codes are capable of reproducing the key features of the problem (Fig. 8). Overall, simulation results are similar in terms of the magnitude of concentrations and the location of fronts. The most significant differences are observed in the region of divergent flow downgradient of the low permeability zone (medium B) near the top of the domain (Fig. 8). In addition, some deviations are observed in the low concentration zone within medium B near the bottom of the domain.

In addition to the solutions computed subject to the CPU time limitation, the participants could also submit solutions using finer meshes without CPU time

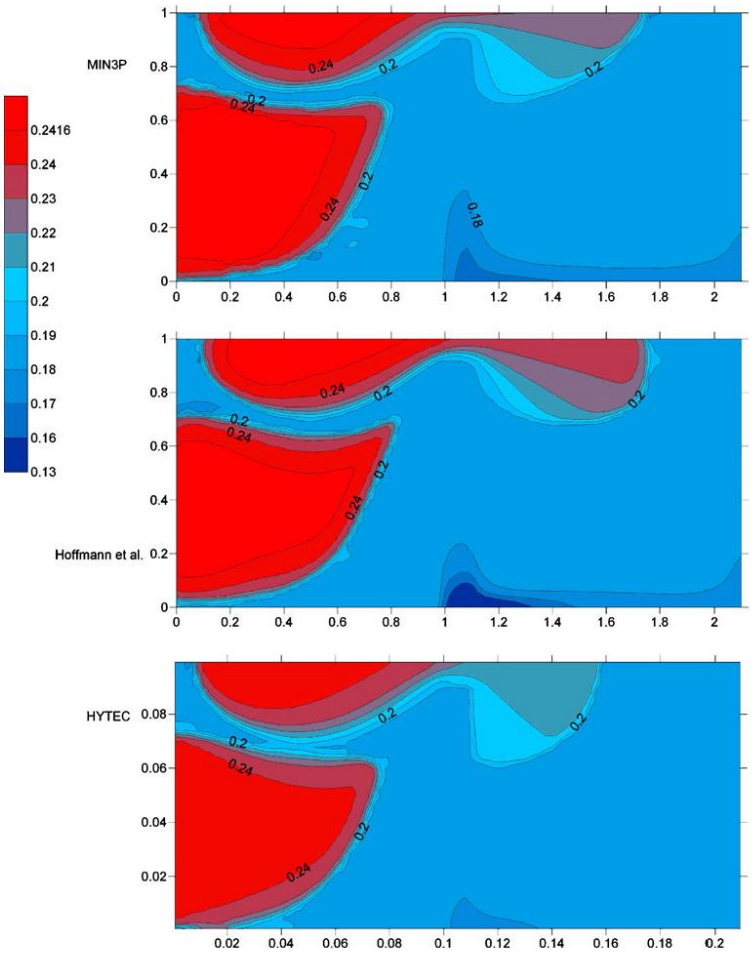


Fig. 8 Concentration contour maps for component X_3 at time 1,000 for the easy 2D advective test case (maximum normalized CPU time is set to 10,000 CPU units)

limitations. In this exercise, Hoffmann et al. used a regular mesh with 107,520 triangles and MIN3P was run with a grid consisting of 21,836 cells (212 × 103). Figure 9 shows the X_3 concentration maps at time 1,000 calculated using these refined meshes. Also, Hoffmann et al. performed a computationally intensive simulation with a 608,256 cells grid, taking 2 weeks on ten processors. The mesh is very fine, and the unstructured mesh used is adapted to describe the meandering flow field. We provide the X_3 concentration map at time 1,000 for this very fine mesh in Fig. 10. The results of the refined simulations show that the grid refinement leads to somewhat sharper concentration fronts and a reduction of local oscillations (Figs. 8, 9, and 10).

However, a more detailed analysis of this aspect was not possible due to the substantial CPU-requirements associated with very fine discretizations. In the time available for this benchmarking project, only the code of Hoffmann et al. was able to compute a solution on such a fine mesh. Hence it was not possible to check this solution with a second code. For this reason, we cannot conclude if the three codes will converge to the same solution and we do not give an error norm because we did not get a reference solution.

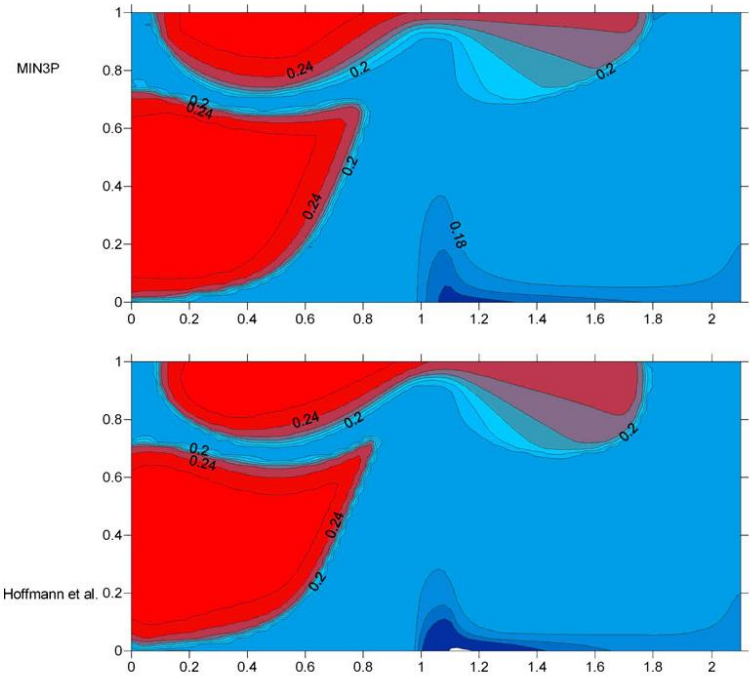


Fig. 9 Concentration contour maps for component X_3 at time 1,000 for the easy 2D advective test case (refined discretization, no CPU time constraint)

5.5 2D easy dispersive test case

The maximum allowed computing time for this case was set to 10,000 CPU units. This benchmark was only completed by two codes. The HYTEC simulation used 840 cells (42 × 20), and MIN3P employed a grid with 5,250 cells (105 × 50), the same discretization as for the 2D advective case.

The results are compared based on the concentration contour map of the immobile component S at time 10 (Fig. 11). S concentrations are depleted completely in the vicinity of the two injection locations, and a very thin and high amplitude S peak appears, similar to the results presented in Figs. 5 and 6 for the 1D easy advective test case. The simulation results from both codes indicate that these narrow and sharp peaks are difficult to resolve in a 2D simulation. A possible remedy would be grid refinement; however, this is difficult to achieve considering the extreme stiffness and high computational demand of this test problem.

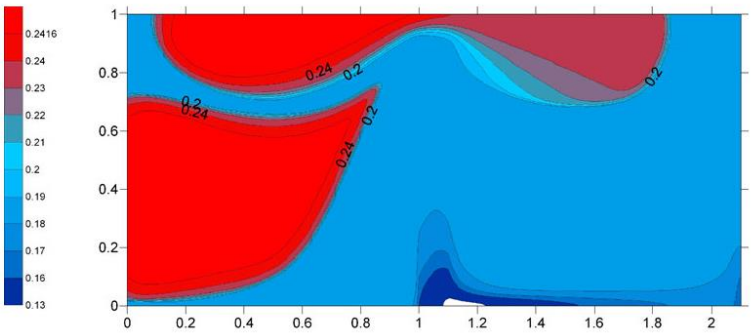


Fig. 10 Concentration contour maps for component X_3 at time 1,000 for the easy 2D advective test case calculated by Hoffmann et al. using a very fine mesh (608,256 cells)

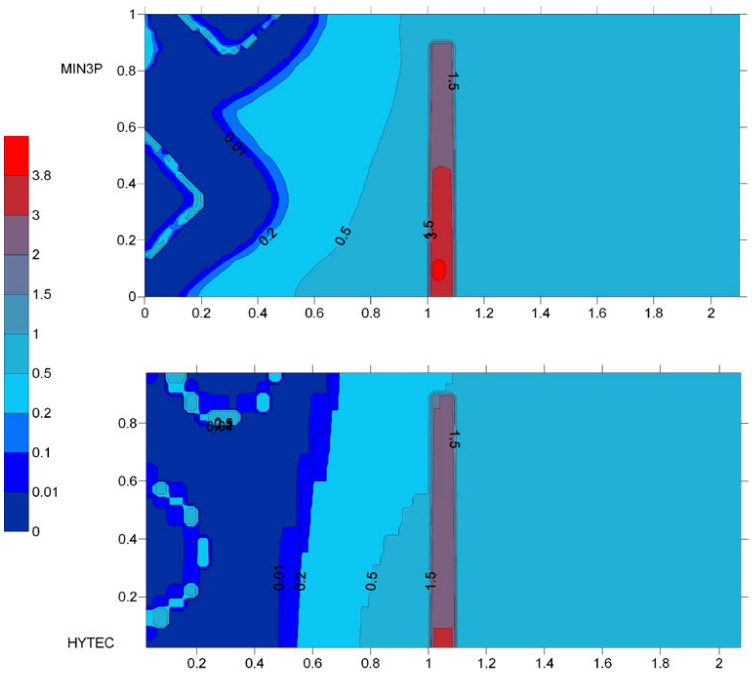


Fig. 11 Concentration contour maps of solid component S at time 10 for the 2D easy dispersive test case

Nevertheless, the results are encouraging in the sense that both simulations produce the same characteristic system behavior.

6 Synthesis of results

6.1 About the benchmark

The staged design of the benchmark was useful because it allowed comparing numerous methods and codes, independent of the level of development. Some of the established codes were able to tackle the benchmark on all three levels, while codes with a more limited reaction network could also participate. Using a fictitious chemical reaction network helped to focus on numerical issues and ensured that differences in the results are due to methods, algorithms, or implementations and not to discrepancies in the geochemistry databases. For the 2D cases, codes with parallel capabilities are needed to solve the problem accurately, i.e., to define a reference solution. Another possibility for future evaluation would be to make the problem "chemically easier" to allow for a quantitative comparison.

6.2 A good confidence in all methods

One of the main outcomes of this benchmark exercise is that the various methods used in this paper for solving reactive transport equations were able to solve the benchmark test cases and to capture their characteristic features both in time and space. Despite some localized differences, the simulation results are quite comparable, which builds confidence in the reactive transport modeling approach in general. Another outcome of this exercise is that some of the codes presented here have been improved to perform this benchmark.

6.3 About sequential approaches

Sequential approaches for reactive transport coupling are attractive because of their highest modularity and flexibility. Since models are becoming increasingly more complex, a modular and "library-based" approach, in which all libraries can be tested as independent modules, is strongly recommended (e.g., as implemented in HYTEC). The sequential approach allows for code development by a team of programmers working relatively independently. Indeed, this method breaks down the reactive transport problem naturally

into three major modules: chemistry, transport, and coupling. Moreover, they allow the use of any chemistry solver with all the knowledge of geochemistry databases. On the other hand, global methods require computing chemistry functions and derivatives and cannot use current chemistry solvers, which do not provide these interfaces. It is well known that operator splitting combined with a noniterative sequential approach (e.g., SPECY) introduces an a priori unknown error. This benchmark illustrates clearly that this method can be used with a rigorous control of errors.

6.4 About global methods

We show with our results that current global approaches can handle large systems describing 1D and 2D reactive transport. As a matter of fact, the simulations of the 2D benchmark were not limited by system memory but by computational time. For the test cases considered, global methods are very competitive in terms of computational efficiency, compared to sequential approaches.

We compared three codes implementing a global approach and using different primary unknowns. Because GDAE1D is based on a differential and algebraic system, it leads to the highest number of coupled unknowns (number of species plus number of components) per number of cells. In a direct substitution approach like in MIN3P, the number of coupled unknowns is reduced to the number of components per number of cells. The reduction scheme implemented by Hoffmann et al. uses even less coupled unknowns, reducing down to three decoupled components per number of cells plus two coupled components per number of cells. A comparison of the CPU time curves (Figs. 2, 3, and 4) illustrates the effect of reducing the number of unknowns. A new version of GDAE1D is under development, where a substitution approach is applied at the linear level. This allows keeping the nice features of DAE solvers with an adaptive time step based on error estimation and an adaptive control of convergence for nonlinear iterations.

6.5 Impact of the dominant transport phenomenon

We show here that all the numerical methods are able to give an accurate solution for both advective and dispersive cases. Nevertheless, it seems that the SNIA method is well adapted for advective problems, with a good tradeoff between accuracy, computational time, and ease of implementation. On the other hand, using the SNIA approach for a dispersive problem must be associated with an increase of the computing cost by reducing the time step or by refining the mesh. The SIA and global approaches are less dependent on the dominant transport phenomenon leading to a good accuracy for both advective and dispersive flows. This accuracy is obtained at the cost of the CPU time for SIA approaches and at the cost of the ease of implementation for global approaches.

6.6 About mesh and time refinement

Looking at Table 2, SPECY is the only code that does not use any adaptive time step. Computing time is lost to perform small time steps during the steady-state period (time between 3,000 and 5,000). An adaptive time step is a very important point to increase the efficiency of a reactive transport code without any loss of accuracy. Nevertheless, all codes compared here use some heuristic methods for time step adaptation based on the convergence rate of the linearization method. Only GDAE1D uses an adaptive order for time discretization and uses an error estimation computed in the DAE solver. This last feature can explain its high accuracy despite the coarse grids used. Further research on reactive transport codes should deal with adaptive time step strategies based on a predictor–corrector scheme or on error estimators.

Looking again at Table 2, some codes use a uniform grid, whereas some other codes refine the mesh in medium B. This mesh refinement reduces significantly computational time. None of the code uses adaptive mesh refinement. This is also a main perspective of research for reactive transport codes.

7 Conclusion and future work

A new benchmark has been designed to compare numerical methods for reactive transport models. This paper presents four different test cases, in 1D and 2D, with advective or dispersive transport conditions. Three classical methods for coupling have been used to solve this benchmark: SNIA with operator splitting (SPECY); SIA (HYTEC), DSA (MIN3P). In addition, two new mathematical methods have been proposed for the solution of reactive transport problems: a DAE approach (GDAE1D) and a reduction scheme (code of Hoffmann et al.). The use of a DAE solver provides an easy way to adapt the time step and to control convergence of Newton iterations, leading to accurate solutions. The reduction scheme presents an important innovation for this field of research, since it allows obtaining accurate solutions at a relatively low computational cost. Implementation of this reduction

scheme may also benefit other approaches. In the case of iterative fixed-point approaches, it could be a way of reducing the number of Picard iterations between chemistry and transport. In the case of noniterative approaches, the reduction method may help to control errors. These two points could be targets for future research.

The most important outcome of this benchmark exercise is that all approaches (SNIA, SIA, DSA, and DAE) were able to generate accurate results for problems of significant complexity and computational difficulty. This finding builds confidence in the use of reactive transport models to help in the assessment of environmental problems in earth sciences and engineering. It has also confirmed that various approaches have different advantages and disadvantages; therefore, a single superior method that is best for all problems cannot be identified. Nevertheless, the good performance of the relatively new code by Hoffmann et al., both in terms of relative accuracy and efficiency, highlights the need for continued collaboration between mathematicians, computer scientists, hydrogeologists, and geochemists.

The benchmark can also be used as a starting point for new comparison exercises. For example, simulations could be enhanced to address a limitation of the current tests. None of the current simulations provide a thorough test for analyzing the effect of transverse dispersion. This deficiency could be removed in the 2D version of the benchmark simply by modifying the boundary conditions to prescribe the injection of different solutions in each injection zone. Dissolved species contained within these solutions would mix along the flowpath and could react with each other subject to either equilibrium or kinetic reactions. In this context, various scenarios could be envisioned, in which the product of the mixing reaction precipitates (equilibrium, kinetically controlled), sorbs, or remains in solution. In addition, the number of components and species could be increased in order to be more representative of real-world reactive transport problems.

Acknowledgements This work has been supported by MoMaS CNRS-2439. We gratefully acknowledge sponsorship of GDR MoMAS by ANDRA, BRGM, CEA, EDF, and IRSN.

References

1. Bain, J.G., Mayer, K.U., Molson, J.W.H., Blowes, D.W., Frind, E.O., Kahnt, R., Jenk, U.: Assessment of the suitability of reactive transport modelling for the evaluation of mine closure options. J. Contam. Hydrol. **52**, 109–135 (2001)
2. Barry, D.A., Miller, C.T., Culligan-Hensley, P.J.: Temporal discretisation errors in non-iterative split-operator approaches to solving chemical reaction/groundwater transport models. J. Contam. Hydrol. **22**, 1–17 (1996)
3. Barry, D.A., Miller, C.T., Culligan, P.J., Bajracharya, K.: Analysis of split operator methods for nonlinear and multispecies groundwater chemical transport models. Math. Comput. Simul. **43**, 331–341 (1997)
4. Bauer, R.D., Rolle, M., Bauer, S., Eberhardt, C., Grathwohl, P., Kolditz, O., Meckenstock, R.U., Griebler, C.: Enhanced biodegradation by hydraulic heterogeneities in petroleum hydrocarbon plumes. J. Contam. Hydrol. **105**, 56–68 (2009)
5. Carnahan, C.L., Remer, J.S.: Nonequilibrium and equilibrium sorption with a linear sorption isotherm during mass transport through an infinite porous medium: some analytical solutions. J. Hydrol. **73**, 227–258 (1984)
6. Carrayrou, J.: Looking for some reference solutions for the reactive transport benchmark of MoMaS with SPECY. Comput. Geosci. (2010, this issue). doi:10.1007/s10596-009-9161-y
7. Carrayrou, J., Mosé, R., Behra, Ph.: A new efficient algorithm for solving thermodynamic chemistry. AIChE. J. **48**, 894–904 (2002)
8. Carrayrou, J., Mosé, R., Behra, Ph.: Modélisation du transport réactif en milieu poreux : schéma itératif associé à une combinaison d'éléments finis discontinus et mixtes-hybrides. Comptes Rendus Ac. Sci Mécanique **331**, 211–216 (2003)
9. Carrayrou, J., Mosé, R., Behra, Ph.: Efficiency of operator splitting procedures for solving reactive transport equation. J. Contam. Hydrol. **68**, 239–268 (2004)
10. Carrayrou, J., Kern, M., Knabner, P.: Reactive transport benchmark of MoMaS. Comput. Geosci. (2010, this issue). doi:10.1007/s10596-009-9157-7
11. de Dieuleveult, C.: Un modèle numérique global et performant pour le couplage géochimie-transport. Ph.D. thesis, University of Rennes 1 (2008)
12. de Dieuleveult, C., Erhel, J.: A global approach to reactive transport: application to the MoMaS benchmark. Comput. Geosci. (2010, this issue). doi:10.1007/s10596-009-9163-9
13. de Dieuleveult, C., Erhel, J., Kern, M.: A global strategy for solving reactive transport equations. J. Comput. Phys. **228**, 6395–6410 (2009)
14. De Windt, L., Burnol, A., Montarnal, P., van der Lee, J.: Intercomparison of reactive transport models applied to UO_2 oxidative dissolution and uranium migration. J. Contam. Hydrol. **61**, 303–312 (2003)
15. De Windt, L., Schneider, H., Ferry, C., Catalette, H., Lagneau, V., Poinssot, C., Poulesquen, A., Jegou, C.: Modeling spent nuclear fuel alteration and radionuclide migration in disposal conditions. Radiochim. Acta **94**, 787–794 (2006)
16. Fahs, M., Carrayrou, J., Younes, A., Ackerer, P.: On the efficiency of the direct substitution approach for reactive transport problems in porous media. Water Air Soil Pollut. **193**, 299–308 (2008)
17. Freedman, V.L., Ibaraki, M.: Coupled reactive mass transport and fluid flow: issues in model verification. Adv. Water Resour. **26**, 117–127 (2003)
18. Henderson, T.H., Mayer, K.U., Parker, B.L., Al, T.A.: Three-dimensional density-dependent flow and multicomponent reactive transport modeling of chlorinated solvent oxidation by potassium permanganate. J. Contam. Hydrol. **106**, 183–199 (2009)
19. Hoffmann, J., Kräutle, S., Knabner, P.: A parallel global-implicit 2-D solver for reactive transport problems in porous

media based on a reduction scheme and its application to the MoMAS benchmark problem. Comput. Geosci. (2010, this issue). doi:10.1007/s10596-009-9173-7
20. Kaluarachchi, J.J., Morshed, J.: Critical assessment of the operator-splitting technique in solving the advection–dispersion–reaction equation: 1. First-order reaction. Adv. Water Resour. **18**, 89–100 (1995)
21. Kanney, J.F., Miller, C.T., Kelley, C.T.: Convergence of iterative split-operator for approximating non-linear reactive transport problem. Adv. Water Resour. **26**, 247–261 (2003)
22. Kräutle, S., Knabner, P.: A new numerical reduction scheme for coupled multicomponent transport–reaction problems in porous media: generalization to problems with heterogeneous equilibrium reactions. Water Resour. Res. **43**, W03429.1–W03429.15 (2007). doi:10.1029/2005WR004465
23. Lagneau, V., van der Lee, J.: HYTEC results of the MoMas reactive transport benchmark. Comput. Geosci. (2010, this issue). doi:10.1007/s10596-009-9159-5
24. Leeming, G.J.S., Mayer, K.U., Simpson, R.B.: Effects of chemical reactions on iterative methods for implicit time stepping. Adv. Water Resour. **22**, 333–347 (1998)
25. Maher, K., Steefel, C.I., White, A.F., Stonestrom, D.A.: The role of reaction affinity and secondary minerals in regulating chemical weathering rates at the Santa Cruz Soil Chronosequence, California. Geochim. Cosmochim. Acta **73**, 2804–2831 (2009)
26. Mayer, K.U., MacQuarrie, K.T.B.: Solution of the MoMaS reactive transport benchmark with MIN3P–model formulation and simulation results. Comput. Geosci. (2010, this issue). doi:10.1007/s10596-009-9158-6
27. Mayer, K.U., Benner, S.G., Frind, E.O., Thornton, S.F., Lerner, D.L.: Reactive transport modeling of processes controlling the distribution and natural attenuation of phenolic compounds in a deep sandstone aquifer. J. Contam. Hydrol. **53**, 341–368 (2001)
28. Mayer, K.U., Frind, E.O., Blowes, D.W.: Multicomponent reactive transport modeling in variably saturated porous media using a generalized formulation for kinetically controlled reactions. Water Resour. Res. **38**, 1174 (2002). doi:10:1029/2001WR000862
29. Mayer, K.U., Benner, S.G., Blowes, D.W.: Process-based reactive transport modeling of a permeable reactive barrier for the treatment of mine drainage. J. Contam. Hydrol. **85**, 195–211 (2006)
30. Molinero, J., Samper, J.: Large-scale modeling of reactive solute transport in fracture zones of granitic bedrocks. J. Contam. Hydrol. **82**, 293–318 (2006)
31. Molins, S., Mayer, K.U.: Coupling between geochemical reactions and multicomponent gas diffusion and advection—a reactive transport modeling study. Water Resour. Res. **43**, W05435 (2007). doi:10.1029/2006WR005206
32. Nowack, B., Mayer, K.U., Oswald, S.E., Van Beinum, W., Appelo, C.A.J., Jacques, D., Seuntjens, P., Gerard, F., Jaillard, B., Schnepf, A., Roose, T.: Verification and intercomparison of reactive transport codes to describe root-uptake. Plant and Soil **285**, 305–321 (2006)
33. Prommer, H., Aziz, L.H., Bolaño, N., Taubald, H., Schüth, C.: Modelling of geochemical and isotopic changes in a column experiment for degradation of TCE by zero-valent iron. J. Contam. Hydrol. **97**, 13–26 (2008)
34. Reeves, H., Kirkner, D.J.: Multicomponent mass transport with homogeneous and heterogeneous chemical reactions: effect of the chemistry on the choice of numerical algorithm. 2. Numerical results. Water Resour. Res. **24**, 1730–1739 (1988)
35. Saaltink, M.W., Carrera, J., Ayora, C.: A comparison of two approaches for reactive transport modelling. J. Geochem. Explor. **69–70**, 97–101 (2000)
36. Saaltink, M.W., Carrera, J., Ayora, C.: On the behavior of approaches to simulate reactive transport. J. Contam. Hydrol. **48**, 213–235 (2001)
37. Salvage, K.M., Yeh, G.T.: Development and application of a numerical model of kinetic and equilibrium microbiological and geochemical reactions (BIOKEMOD). J. Hydrol. **209**, 27–52 (1998)
38. Selim, H.M., Mansell, R.S.: Analytical solution of the equation for transport of reactive solutes through soils. Water Resour. Res. **12**, 528–532 (1976)
39. Shen, H., Nikolaidis, N.P.: A direct substitution method for multicomponent solute transport in ground water. Ground Water **35**, 67–78 (1997)
40. Siegel, P., Mosé, R., Ackerer, Ph., Jaffre, J.: Solution of the advection–diffusion equation using a combination of discontinuous and mixed finite elements. Int. J. Num. Methods Fluids **24**, 595–613 (1997)
41. Spiessl, S.M., MacQuarrie, K.T.B., Mayer, K.U.: Identification of key parameters controlling dissolved oxygen migration and attenuation in fractured crystalline rocks. J. Contam. Hydrol. **95**, 141–153 (2008)
42. Steefel, C.I., Lasaga, A.C.: A coupled model for transport of multiple chemical species and kinetic precipitation/dissolution reactions with application to reactive flow in single phase hydrothermal systems. Am. J. Sci. **294**, 529–592 (1994)
43. Steefel, C.I., MacQuarrie, K.T.B.: Approaches to modelling of reactive transport in porous media. In: Lichtner, P.C., Steefel, C.I., Oelkers, E.H. (eds.) Reactive Transport in Porous Media, vol. 34, pp. 82–129. Reviews in Mineralogy, Mineralogical Society of America, Washington (1996)
44. Steefel, C.I., Carroll, S., Zhao, P.H., Roberts, S.: Cesium migration in Hanford sediment: a multisite cation exchange model based on laboratory transport experiments. J. Contam. Hydrol. **67**, 219–246 (2003)
45. Sun, Y., Petersen, J.N., Clement, T.P.: Analytical solutions for multiple species reactive transport in multiple dimensions. J. Contam. Hydrol. **35**, 429–440 (1999)
46. Toride, N., Leij, F.J., van Genuchten, M.T.: A comprehensive set of analytical solutions for nonequilibrium solute transport with first-order decay and zero-order production. Water Resour. Res. **29**, 2167–2182 (1993)
47. Valocchi, A.J., Malmstead, M.: Accuracy of operator-splitting for advection–dispersion–reaction problems. Water Resour. Res. **28**, 1471–1476 (1992)
48. van Genuchten, M.T.: Analytical solutions for chemical transport with simultaneous adsorption, zero-order production and first-order decay. J. Hydrol. **49**, 213–233 (1981)
49. van Genuchten, M.T., Wierenga, P.J.: Mass transfer studies in sorbing porous media. 1. Analytical solutions. Soil Sci. Soc. Am. J. **40**, 473–480 (1976)
50. van Genuchten, M.T., Wierenga, P.J., O'Connor, G.A.: Mass transfer studies in sorbing porous media. 3. Experimental evaluation with 2,4,5-T[1]. Soil Sci. Soc. Am. J. **41**, 278–285 (1976)
51. van der Lee, J., De Windt, L., Lagneau, V., Goblet, P.: Module-oriented modeling of reactive transport with HYTEC. Comput. Geosci. **29**, 265–275 (2003)
52. van der Lee, J., Langeau, V.: Rigorous methods for reactive transport in unsaturated porous medium coupled with chemistry and variable porosity. In: Miller, C.T.,

Farthing, M.W., Gray, W.G., Pinder, G.F. (eds.) Computational methods in water resources (CMWR XV), vol. 48(1), pp. 861–868. Elsevier (2004)
53. Walter, A.L., Frind, E.O., Blowes, D.W., Ptacek, C.J., Molson, J.W.: Modelling of multicomponent reactive transport in groundwater, 2. Metal mobility in aquifers impacted by acidic mine tailings discharge. Water Resour. Res. **30**, 3149–3158 (1994)
54. Yeh, G.T., Tripathi, V.S.: A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components. Water Resour. Res. **25**, 93–108 (1989)

Annexe 8. Parameter estimation for reactive transport by a Monte-Carlo approach

Parameter Estimation for Reactive Transport by a Monte-Carlo Approach

Mohit Aggarwal and Jérôme Carrayrou
Institut de Mécanique des Fluides et des Solides de l'Université Louis Pasteur, UMR 7507 Université Louis Pasteur-CNRS, 2 rue Boussingault, 67000 Strasbourg, France

DOI 10.1002/aic.10813
Published online April 17, 2006 in Wiley InterScience (www.interscience.wiley.com).

The chemical parameters used in reactive transport models are not known accurately due to the complexity and the heterogeneous conditions of a real domain. The development of an efficient algorithm in order to estimate the chemical parameters using Monte-Carlo method is presented. By fitting the results obtained from the model with the experimental curves obtained with various experimental conditions, the problem of parameters estimation is converted into a minimization problem. Monte-Carlo methods are very robust for the optimization of the highly nonlinear mathematical model describing reactive transport. It involves generating random values of parameters and finding the best set. The focus is to develop an optimization algorithm which uses less number of realizations so as to reduce the CPU time. Reactive transport of TBT through natural quartz sand at seven different pHs is taken as the test case. Our algorithm will be used to estimate the chemical parameters of the sorption of TBT onto the natural quartz sand. © 2006 American Institute of Chemical Engineers *AIChE J*, 52: 2281–2289, 2006
Keywords: reactive transport, numerical modeling, Monte-Carlo simulation, parameters estimation

Introduction

Recent developments in reactive transport modeling makes reactive transport models more and more used for simulating, describing, improving comprehension, or providing prevision[1] for hydrogeologists or geochemists. They are used in highly sensitive domains, such as ground water supply preservation, development of pollution remediation scenarii or nuclear waste storage study. Geochemical systems described by these numerical models become more and more complex, mainly by an increase of the chemical phenomenon complexity. A reactive transport model needs five kinds of inputs: the hydraulic properties of the domain, the chemical reactions occurring and the chemical parameters, the boundary and initial conditions. We will focus here on the chemical part of the problem. To describe a reactive transport model, a set of chemical reactions is first chosen and after that the corresponding parameters are obtained, whatever the way. Nevertheless, the required chemical parameters are not known exactly. Because the described phenomena are nonlinear, low precision on the determination of the parameters can lead to rejection of an accurate set of reactions. Moreover, it is impossible to determine if the divergence between the experimental results, and the calculated one is due to a false set of reactions or insufficiently precise determination of the parameters. Today, parameters estimation is mainly done using batch experiments with unique experimental conditions with tools like FITEQL.[2] Estimated parameters are then accurate for batch reactor and for a given experimental state, such as imposed pH or fixed ionic force. Extrapolating these parameters to natural uncontrolled systems is then very hazardous. The aim of this work is to present a way to estimate parameters for most realistic cases by a Monte-Carlo procedure. The reactive transport parameters will be estimated from column experiments. Calculated elution curves of an injection-leaching experiment will be fitted to experimental elution curves. In order to lead to parameters that will not (or less) depend on the experimental conditions; many elution curves obtained at different pH will be simultaneously fitted.

Because of the high nonlinearity of the mathematical model

Correspondence concerning this article should be addressed to J. Carrayrou at carrayro@imfs.u-strasbg.fr
Mohit Aggarwal is also affiliated with Indian Institute of Technology, Dept. of Chemical Engineering, New Dehli, India

describing reactive transport under instantaneous equilibrium assumption, we prefer a Monte-Carlo method instead of a gradient-like one. Indeed, it is well known that gradient-like methods are very efficient if there is a linear relation between the parameters and the objective function. Nonlinear equations are then always associated to difficulties on performing inverse modeling as explain by Zhang and Guay.[3] For nonlinear relations, such as equilibrium chemical equations, their efficiency decreases strongly. It has been well established[4,5] that gradient methods are sometime unable to solve some batch equilibrium problems. For batch equilibrium, nonconvergence of gradient method is due to local minima, flat zone of the error function or infinite loop phenomena.[5] Extrapolation of these conclusions to inverse modelling of reactive transport lets expect many convergence problems for gradient like methods as assumed by Marshall.[6] Even if the Monte-Carlo methods are more time-consuming, they are indeed more robust than the gradient-like methods because they do not use the slope of the objective function. A Monte-Carlo method has been successfully used for sensitivity analysis of coagulation processes.[7] Moreover, these authors used a Monte-Carlo approach to obtain the derivative of the objective function before running a gradient-like method for the parameters estimation.

We first present the numerical model: the advection-dispersion-reaction (ADR) equation is detailed, and we present the methods used to solve it as fast as possible. After that, we explain the methods used for parameter estimations: the Monte-Carlo procedure, the error function and the optimization procedures.

A results section is devoted to the study of the developed algorithm. A column experiment on sorption of tributyltin onto a natural quartz sand[8,9] at seven different pH is used as test case. We show the efficiency of the algorithm developed. We then discuss the results obtained through our optimization procedure, and we compare them to results obtained from batch calculation at one pH.

Numerical Model

Reactive transport equation

Under the assumptions of instantaneous equilibrium and identical dispersion of solutes, the ADR equation can be written like Eq. 1 for the Nx component j of the system

$$\omega \frac{\partial (Td_j + Tf_j)}{\partial t} = \nabla \cdot [\mathbf{D} \cdot \nabla (Td_j)] - U \cdot \nabla (Td_j) \quad \text{for } j = 1 \text{ to } Nx \quad (1)$$

where, ω (−) is the active porosity, $\mathbf{D}$ (m^2/s) the dispersion tensor, U (m/s) the Darcy velocity, Td_j (mol/dm^3) the total mobile concentration of component j and Tf_j (mol/dm^3) the total nonmobile concentration of component j. The chemical system is described by mass action and conservation laws. The mass action laws Eq. 2 are written for the formation of the Nc species C_i by the selected component set X_j

$$\{C_i\} = K_i \cdot \prod_{j=1}^{Nx} \{X_j\}^{a_{i,j}} \quad \text{for } i = 1 \text{ to } Nc \quad (2)$$

where the activity of species and component is noted $\{-\}$, K_i is the equilibrium constant, Nx the number of components used to describe the system, and $a_{i,j}$ the stochiometric coefficient for the mass action law. For precipitated minerals, the mass action laws are written under the precipitation product form Eq. 3 if precipitation occurs

$$1 = Kp_i \cdot \prod_{j=1}^{Nx} \{X_j\}^{ap_{i,j}} \quad (3)$$

Kp_i is the precipitation product of precipitated species Cp_i and $ap_{i,j}$ are the stochiometric coefficients. The conservation law Eq. 4 is written to conserve the total quantity $[T_j]$ (mol/dm^3) of each component

$$[T_j] = \sum_{i=1}^{Nc} b_{i,j} \cdot [C_i] + \sum_{i=1}^{NcP} bp_{i,j} \cdot [Cp_i] \text{ for } j = 1 \text{ to } Nx \quad (4)$$

where the concentrations (mol/dm^3) are noted $[-]$, Nc (resp., NcP) is the number of species (resp., precipitated) in the system, and $b_{i,j}$ (resp., $bp_{i,j}$) the stochiometric coefficient of species C_i (resp. Cp_i) for the conservation law. Stochiometric coefficients used for the mass-action law are different than those used for the conservation law in the model SPECY.[10] This allows the description of some geochemical processes, such as surface precipitation.[11]

Substituting the species activity from Eq. 2 instead of the species concentration into the conservation law Eq. 4 is done by using activity coefficient γ_i and leads to the nonlinear algebraic system Eq. 5.

$$[T_j] = \sum_{i=1}^{Nc} b_{i,j} \cdot \frac{K_i}{\gamma_i} \prod_{k=1}^{Nx} (\gamma_k [X_k])^{a_{i,k}} + \sum_{i=1}^{NcP} b_{i,j} \cdot [Cp_i] \quad \text{for } j = 1 \text{ to } Nx$$

$$1 = Kp_i \cdot \prod_{j=1}^{Nx} \{X_j\}^{ap_{i,j}} \text{ for } i = 1 \text{ to } NcP \quad (5)$$

The system Eq. 5 is of size $Nx + NcP$, and the unknowns are the component $[X_j]$, and the precipitated species $[Cp_i]$ concentrations (mol/dm^3). Activity coefficient are calculated using the Davies model,[11,12,13] which is accurate for ionic strength $I < 0.5$

$$\log(\gamma_i) = -A \cdot z_i^2 \cdot \left(\frac{\sqrt{I}}{1 + \sqrt{I}} - B \cdot I \right) \quad (6)$$

where $A = 1.82 \cdot 10^6 \cdot (\varepsilon \cdot T)^{-3/2}$, T is the Kelvin temperature, ε the electric permeability of water and $B = 0.24$.

The mobile total concentration Td_j, and the nonmobile total concentration Tf_j of component j are calculated from the sum of the concentration of the NcD mobile species, and NcF non-mobile species

$$[Td_j] = \sum_{i=1}^{NcD} b_{i,j} \cdot [C_i] \text{ and } [Tf_j] = \sum_{i=1}^{NcF} b_{i,j} \cdot [C_i] + \sum_{i=1}^{NcP} b_{i,j} \cdot [Cp_i] \quad \text{for } j = 1 \text{ to } Nx \quad (7)$$

Because the solution of Eq. 5 is unique for a given set of T_j, the combination of Eq. 5 and Eq. 7 can be noted Eq. 8

$$[Td_j] = fd_j([T_j]) \text{ and } [Tf_j] = ff_j([T_j]) \text{ for } j = 1 \text{ to } Nx \quad (8)$$

where the total concentration T_j is the sum of the mobile Td_j, and nonmobile Tf_j total concentration for each component.

Sorption phenomena can be described easily by ion exchange or by surface complexation. For ion exchange, the mass action law describing the formation of a species is given in Eq. 2. For surface complexation phenomenon, the sorption site should be defined as a component X_s. Then the potential of the surface Ψ is added to the mass action law describing the sorption of a species Cs_i

$$\{Cs_i\} = K_i \cdot \exp\left(-\frac{z_i F}{R\tau} \cdot \Psi\right) \cdot \prod_{j=1}^{Nx} \{X_j\}^{a_{i,j}} \quad (9)$$

where z_i is the charge of the species Cs_i, R is the gas constant, F is the Faraday constant, and τ is the temperature.

Different models can be used to obtain the potential Ψ from the electrostatic charge fixed at the surface.[11,14,15] In this work, only the diffuse layer model (DLM) will be used

$$\sum_{sorbed} z_i \cdot [Cs_i] = \frac{S \cdot M}{F} (8 \cdot R \cdot \tau \cdot \varepsilon \cdot \varepsilon_0 \cdot I)^{1/2} \cdot \sinh\left(\frac{Z_{el} \cdot F \cdot \Psi}{2 \cdot R \cdot \tau}\right) \quad (10)$$

where ε_0 is the permittivity of vacuum, ε the permittivity of water, Z_{el} the electrical charge of counterion, S the specific area of the solid, and M the mass concentration of the solid. By defining the electrostatic potential as a component X_Ψ and the associated stochiometric coefficient $a_{i,\Psi}$

$$\{X_\Psi\} = \exp\left(-\frac{F}{R\tau} \cdot \Psi\right) \text{ and } a_{i,\Psi} = z_i \quad (11)$$

We can include the complexation surface phenomena into the general formulation presented by Eqs. 2 and 4. The mass action law Eq. 9 is also written

$$\{Cs_i\} = K_i \cdot \{X_\Psi\}^{a_{i,\Psi}} \cdot \prod_{j=1}^{Nx} \{X_j\}^{a_{i,j}} \quad (12)$$

and the conservation law Eq. 4 is expressed as Eq. 13 with the DLM model Eq. 10.

$$T_\Psi = \frac{S \cdot M}{2F} \cdot \sqrt{8 \cdot R\tau \cdot \varepsilon \cdot \varepsilon_0 \cdot I} \cdot (\{X_\Psi\}^{-(Z_{el}/2)} - \{X_\Psi\}^{Z_{el}/2}) = \sum_{sorbed} a_{i,\Psi} \cdot [Cs_i] \quad (13)$$

Reactive transport model

In order to reduce to minimum the computation time, efficient numerical methods are used. The computer code SPECY solves the advection-dispersion-reaction Eq. 1 by an operator-splitting scheme. As shown by several authors[10,16,17] the best way to solve ADR equation under instantaneous equilibrium assumption by OS is to use a standard iterative scheme.[18]

The resolution procedure is described by Eqs. 14 and 15

$$\omega \frac{Td_j^{n+1,k+1} - Td_j^n}{\Delta t} = \nabla \cdot [\mathbf{D} \cdot \nabla(Td_j)] - U \cdot \nabla(Td_j) - \omega \frac{Tf_j^{n+1,k} - Tf_j^n}{\Delta t} \text{ for } j = 1 \text{ to } Nx \quad (14)$$

$$[Tf_j^{n+1,k+1}] = ff_j([Td_j^{n+1,k+1}] + [Tf_j^{n+1,k}]) \quad \text{for } j = 1 \text{ to } Nx, \text{ with } [Tf_j^{n+1,0}] = [Tf_j^n] \quad (15)$$

The advective part of the equation is solved by the discontinuous finite element method and the dispersive part by the mixed hybrid finite element method. The combination of this resolution method and a standard iterative OS scheme leads to a very accurate solution, even if the mesh size is large.[18] It is well known that the maximum computing time is spent for the geochemical computation and not for the transport one. To reduce geochemical computation, SPECY solves the nonlinear algebraic system Eq. 5 with an efficient algorithm.[5] The positive continuous fraction method is used as a preconditioner to obtain an intermediate solution, close to the exact one. Then the Newton-Raphson method is used to obtain the final solution. To increase the efficiency of the Newton-Raphson method, SPECY limits the research procedure to the chemically allowed interval. This specific algorithm reduces the computing time of geochemical computation.

All these implementations lead to faster computing of reactive transport phenomena. These implementations are useful because the number of realizations needed by a Monte-Carlo approach is very high.

Monte-Carlo procedure

In order to reduce the computing time, the number of parameters N_P which will be estimated is reduced to a minimum. Hydrodynamic parameters, such as porosity, velocity or dispersivity are estimated from tracer experiment, and are assumed to be well known. By the same way, chemical parameters for aqueous reactions, such as equilibrium constants and concentrations are obtained from thermodynamic databases, and are assumed to be well known. Only chemical parameters related to the solid-water interface phenomena are estimated, that is, equilibrium constant for sorption reactions, total concentration for sorption components and surface complexation parameter, such as specific area or capacitance.

Parameters **P** which have to be estimated are randomly generated. The distribution of each parameter P_i describes a Gaussian distribution. Each Gaussian distribution is set by its mean $\bar{P}_i$ and its standard deviation σ_i. There are two ways of sampling random numbers from a given set of probability distribution:

(1) Monte Carlo (MC): In this method we generate a random probability value between the given range and map it onto the corresponding parameter value. To generate more parameter values, we just repeat the procedure independently. Using this method, we need large sets of parameter values to have a distribution close to the actual one.

(2) Latin Hyper Cube (LHC): In LHC we divide the given range of parameter values into regions of equal probability. A random value of parameter is then generated in each interval. This method ensures that we cover the whole range of parameter values, and it gives distribution close to the actual one even if the number of generated random values is limited.

It is clear that the LHC method is more accurate and efficient.[19] Thus, we will be using LHC method to generate random numbers.

Objective function

In order to compare the generated set of parameter **P** and to find the best one, the following objective function F_D is used

$$F_N(\mathbf{P}) = \frac{1}{N_{Exp}} \sum_{Experiment} \left(\frac{\theta_{Exp}}{N_{Mes}} \cdot \sum_{Mesure} \theta_{Mes} \sqrt{\left[\frac{C_{Mes} - C_{Calc}(\mathbf{P})}{C_{Mes} + \varepsilon} \right]^2} \right) \tag{16}$$

where N_{Exp} is the number of different experiments to fit, θ_{Exp} is the weight of each experiment. N_{Mes} is the number of experimental points for each experiment, θ_{Mes} is the weight of each experimental point, C_{Mes} is the measured concentration, and C_{Calc} is the calculated concentration. The weight of each experiment θ_{Exp} and of each experimental point θ_{Mes} are defined by the modeller in order to give more or less importance to an experiment or to a point. These values are generally defined depending on the measured error. The limiting parameter ε is used to reduce the influence of very small concentration in the construction of the objective function. Practically, we set ε equal to the detection limit of measurement procedure used during the experiments. We use $\frac{C_{Mes} - C_{Calc}}{C_{Mes} + \varepsilon}$ in the objective function in order to give the same importance to small concentrations than to higher one.

Optimization procedure

The objective of the optimization procedure is to find the best set of parameters. It is well known that Monte-Carlo procedures are time consuming. In order to reduce CPU time, we combine an automatic procedure to find a minimum of the objective function and an expert analysis to determine if the given minimum is a local or a global one. We have tested other procedures, such as simulated annealing but these kinds of methods are too much time consuming in our case. Because the computing time of one realization is quite long, it is more efficient to analyze manually the validity of the proposed set of parameters.

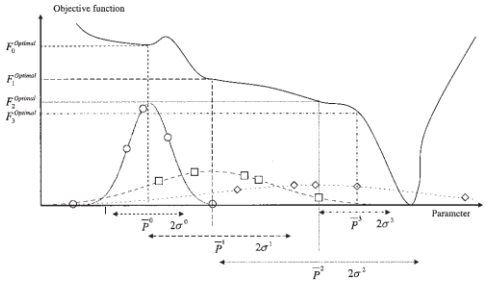


Figure 1. Optimization procedure for a single parameter problem.

We plot the Gaussian curves used for the parameter generator and $N_I = 5$ parameter values for three optimization steps.

Automatic optimization. To reduce the CPU time, we change the mean $\bar{\mathbf{P}}$ and the standard deviation $\boldsymbol{\sigma}$ of the Gaussian curves used to generate the parameters **P** during the optimization procedure.

We first generate $N_I = 100$ sets of parameters $\mathbf{P}^k$ from $\bar{\mathbf{P}}^i$ and $\boldsymbol{\sigma}^i$ (see Figure 1). Initial mean $\bar{\mathbf{P}}^0$ and standard deviation $\boldsymbol{\sigma}^0$ are given by the modeler. The first minimal value of the objective function is $F^{Min} = F(\bar{\mathbf{P}}^0)$. The reactive transport problem is solved with these sets of parameters, and the objective function $F^k(\mathbf{P}^k)$ is then calculated for $k = 1$ to N_I. We then compare the N_I objective functions and select the smallest $F^{Optimal}$. If $F^{Optimal} < F^{Min}$ then we change the mean by taking $\bar{\mathbf{P}}^{i+1} = \mathbf{P}^{Optimal}$, the standard deviation by taking

$$\boldsymbol{\sigma}^{i+1} = \mathrm{Max}[|\mathbf{P}^{Optimal} - \bar{\mathbf{P}}^i|; (\alpha_\sigma \cdot \boldsymbol{\sigma}^i)]. \tag{17}$$

and the value of F^{Min} by taking $F^{Min} = F^{Optimal}$ as shown in Figure 1.

We use Eq. 17 to adapt the standard deviation. α_σ is a parameter which have to limit the reduction of the standard deviation. The optimal set of parameters $\mathbf{P}^{Optimal}$ is sometimes very close to the previous one $\bar{\mathbf{P}}^i$. This can lead to a very fast reduction of the standard deviation and, consequently, a severe reduction of the convergence speed of the algorithm. In this work, we use $\alpha_\sigma = 0.5$ as shown in Figure 2. This ensures a sufficiently slow reduction of the standard deviation and a good convergence speed.

If $F^{Optimal} > F^{Min}$, we reject the proposed set of parameters and increase the value of N_I by taking $N_I = 1{,}000$. Then we run again N_I realizations with the same mean $\mathbf{P}^i$ and $\boldsymbol{\sigma}^i$. In this case, we change the mean and the standard deviation immediately when a better set is found. After that, if no minimum is found, the standard deviation is decreased and $\boldsymbol{\sigma}^{i+1} = \boldsymbol{\sigma}^i/2$. We have tested to increase the standard deviation. However, this choice requires more realizations to cover all the ranges of the parameter values with a sufficiently small interval, and is very CPU time consuming. It seems that the minimum of the error function is a very small hole in a quite flat area, as for batch equilibrium calculation.

We choose $N_I = 100$ at the beginning of the algorithm

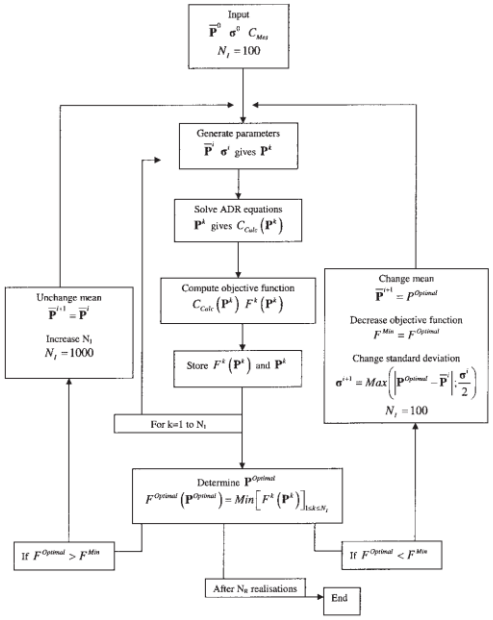


Figure 2. Optimization algorithm.

because this number of realizations is enough to give a good representation of a Gaussian curve for each parameter using the LHC sampling method. Decreasing the value of N_I reduces the robustness of the algorithm, but it increases the convergence rate. Convergence becomes then less and less probable. Increasing the value of N_I increases the robustness of the algorithm, making convergence to the accurate solution more and more certain, but increases the computing time too.

This cycle (running N_I realization, selecting the best set, changing the mean and the standard deviation) is usually done 20 times. The total number of realizations is then $N_R = 2{,}000$. The best set of parameters and the associated standard deviation have to be analyzed to determine their validity.

Expert analysis

The total number of realization used here ($N_R = 2{,}000$) is not large enough to ensure the convergence to the global minima. Some numerical methods can be used to increase the probability of convergence: increasing N_R; performing simulated annealing. Unfortunately, these methods are too time consuming in our case. In order to reduce the time needed to find the minimum, we use the following procedure:

(a) Running three to six automatic optimizations using the initial guess of parameters.

(b) Selecting one to three acceptable sets of parameters.

(c) Running one to three automatic optimizations using each selected set of parameters. The initial standard deviations can be smaller than those used at the stage (a).

(d) Repeating stages (b) and (c) until there is only one set of parameters left. The initial standard deviations can be reduced. This set of parameters cannot be improved by an automatic optimization.

(e) Analyzing the correlations between the fitted parameters.

This procedure ensures that the fitted set of parameters correspond to a minimum of the objective function.

The selection of the acceptable sets of parameters (stage b) uses the objective functions and some additional informations. The sets associated to the higher objective functions are rejected. It is useful to have some additional informations which are not injected into the optimization procedures. These informations can be estimated values of some parameters, or a relation between some parameters. They can be obtained by another experiment than those used to perform the optimization.

Results and Discussion

Reactive transport test-case

The reactive transport test-case we use has been given by Bueno et al.[8,9] It is about transport of TBT through a natural quartz sand. These authors provide breakthrough curves of TBT at seven different pH.[8] This leads to seven different sets of Langmuir parameters. From Langmuir parameters obtained by these authors at pH = 6.1, equilibrium constants and site concentration for the sorption of TBT onto the natural quartz sand can be calculated. Specific area of this sand is given by Bueno et al.[8] All these parameters and the chemical reactions assumed in the system are summarized in Table 1.

Table 1. Morel Tableau for the TBT Reactive Transport Test-Case

	H^+	Cl^-	NO_3^-	Na^+	Im	TBT^+	≡S—OH	Ψ_S	Given log (K)
OH^-	−1								−14.0
Im H^+	1				1				7.0
Im TBT^+					1	1			3.91
TBTOH	−1					1			−6.25
TBTCl		1				1			0.6
$TBTNO_3$			1			1			0.62
≡S—OH_2^+	1						1	1	**4**
≡S—O^-	−1						1	−1	**−8**
≡S—OTBT	−1					1	1		**1.37**
≡S—$OHTBT^+$						1	1	1	**5.46**
≡S—ONa	−1			1			1		**−5.3**
Initial condition (M)	Fixed	0.0	0.1	0.1	10^{-3}	0.0	$\mathbf{10^{-5}}$		
Injection (M)	Fixed	$8.6\ 10^{-6}$	0.1	0.1	10^{-3}	$8.6\ 10^{-6}$			
Leaching (M)	Fixed	0.0	0.1	0.1	10^{-3}	0.0			

Given specific area[9] S = **0.200** $m^2 \cdot g^{-1}$. Bold parameters will be estimated.

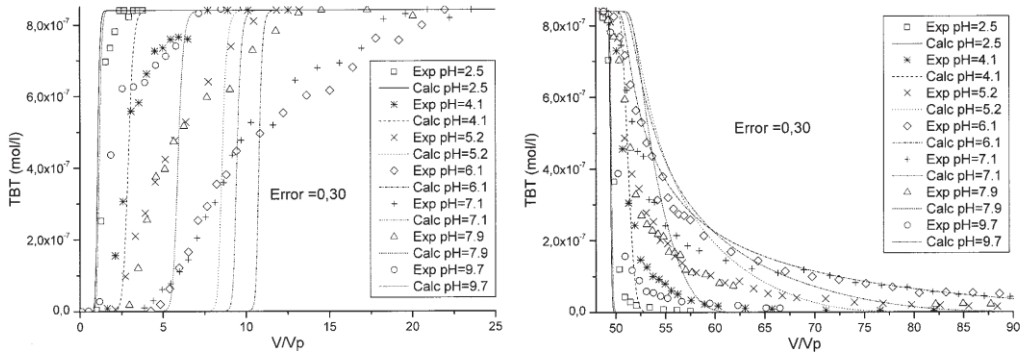


Figure 3. Experimental and initially calculated elution curves.

Initially calculated elution curves are obtained with *batch initial* parameters from Bueno et al.[8] Error function and parameters values are given in Table 2a and Figure 3a: injection of TBT into the column. Figure 3b leaching of TBT out of the column.

A measure of the point of zero charge (PZC) of the sand has been done.[9] The PZC of a surface is the pH where this surface is electrically neutral in pure water. Table 1 shows that in pure water, the quartz sand is supposed to form two species: $\equiv S\!-\!OH_2^+$ and $\equiv S\!-\!O^-$. The PZC is used here as additional information to select the acceptable sets of parameters by controlling the relation between the two equilibrium constants: $K_{\equiv S-OH_2^+}$ and $K\equiv S\!-\!O^-$. The experimental measure[9] gives

$$PZC = \frac{1}{2}\left[\log(K_{\equiv S-OH_2^+}) - \log(K_{\equiv S-O^-})\right] = 6 \pm 1 \quad (18)$$

The experimental column is 20 cm long, discretized in 20 cells. We use a Courant number equal to one, and a Pèclet number equal to five

$$CFL = \frac{U\Delta t}{\Delta x} = 1 \text{ and } Pe = \frac{U\Delta x}{D} = 5 \quad (19)$$

On the contrary of standard numerical methods (finite volume or finite element) which need a Peclet number less than two, the association of discontinuous finite element and mixed hybrid finite element used here has been successfully tested[20] for Peclet numbers less to 100.

Breakthrough curves obtained with these given parameters are compared with experimental ones in Figure 3. Elution part of the pH = 6.1 breakthrough curve is very well fitted to experimental data (Figure 3b). Indeed, we used parameters estimated at this pH. Unfortunately, the fitting of the elution curves to experimental data is not so good for other pHs, and not any injection curves is well described. The modelized system leads to sharp injection fronts instead of the experimental one which are a little diffusive (Figure 3a). In this work, we will answer the question: "Is the proposed set of parameters inaccurate, or is the set of reactions which does not describe accurately the phenomena?"

We will test the optimization procedure over this case in order to obtain one set of chemical parameter describing simultaneously these seven experiments. We set the same weights to all seven experiments, so that $\theta_{\text{Exp}} = 1$. There are a lot of experimental points at small concentrations (less than 10^{-7} mol · l^{-1}). Small concentrations are more represented, and then have a more important influence on the optimization procedure than higher concentrations. In order to give the same influence to all the concentration ranges we set the weight of experimental points as

$$\text{If } C_{Mes} > 10^{-7} \text{ mol} \cdot l^{-1} \text{ then } \theta_{Mes} = 1 \text{ else } \theta_{Mes} = 0.2 \quad (20)$$

The implementations presented previously to reduce CPU time allow us to run 2,000 realizations of these seven reactive transport problems in 26 h, on a COMPAQ Professional Workstation XP1000, with a physical memory of 1280.00 megabytes, and a 500 mHz EV6 processor.

Parameters estimation

The parameters estimation began from the *batch initial* set of parameters, with a restricted standard deviation (see Table 2a). We run $N_R = 2{,}000$ realizations, and $N_I = 100$. Through parameters estimation, the objective function is reduced from $F_N = 0.30$ for the *batch initial* parameters to $F_N = 0.261$ for the worst optimized set (Table 2b), and $F_N = 0.192$ for the best one (Table 2e). The set of parameters given in Table 2c is representative of the set usually found. As presented in the "expert analysis" section, we performed five searches from the same *batch initial* set.

We give in Table 2b the worst set found. Many indicators can help us to reject this set of parameters. The residual standard deviations are still large, more than 0.2 for equilibrium constants. Moreover, the *PZC* of the natural quartz sand is not accurately given by the estimated parameters

$$PZC = \frac{1}{2}\left[\log(K_{\equiv S-OH_2^+}) - \log(K_{\equiv S-O^-})\right] = \frac{1}{2}(6.17 + 8.04) = 7.10 \quad (21)$$

Table 2. Results of Parameter Estimation

F_N	Batch Parameters Figure 3 (a)		Worth Set Found Figure 4 (b)		Set Usually Found (c)		Best Set First Found (d)		2nd Optimization from (d) Figure 4 (e)		Longer Research N_R = 20 000 N_I = 1 000 (f)	
	0.30		0.261		0.204		0.201		0.1921		0.195	
	$\bar{P}$	σ	$\bar{P}$	σ	$\bar{P}$	σ	$\bar{P}$	σ	$\bar{P}$	σ	$\bar{P}$	σ
$\log(K_{\equiv S-OH_2^+})$	4	4	6.17	0.27	1.82	0.58	2.25	$3.8\ 10^{-2}$	2.90	$3.8\ 10^{-2}$	3	$2.1\ 10^{-2}$
$\log(K_{\equiv S-O^-})$	−8	4	−8.04	1.77	−9.35	0.31	−10.4	$2.3\ 10^{-2}$	−7.58	$3.8\ 10^{-2}$	−7.7	$4.3\ 10^{-2}$
$\log(K_{\equiv S-OTBT})$	1.37	4	0.99	0.43	1.46	$7.0\ 10^{-2}$	1.45	$5.4\ 10^{-3}$	1.16	$1.0\ 10^{-2}$	0.99	$9.2\ 10^{-3}$
$\log(K_{\equiv S-OHTBT^+})$	5.46	4	7.89	0.32	5.17	$7.3\ 10^{-2}$	5.30	$4.5\ 10^{-3}$	5.62	$4.0\ 10^{-3}$	5.7	$6.9\ 10^{-3}$
$\log(K_{\equiv S-ONa})$	−5.3	4	−7.03	0.45	−6.23	$4.1\ 10^{-2}$	−6.27	$3.9\ 10^{-3}$	−7.46	$3.2\ 10^{-1}$	−8.8	$8.9\ 10^{-2}$
$[\equiv S-OH]$ (mol/l)	10^{-5}	10^{-5}	$5.78\ 10^{-6}$	$3.9\ 10^{-7}$	$6.55\ 10^{-6}$	$4.9\ 10^{-7}$	$6.44\ 10^{-6}$	$3.0\ 10^{-8}$	$6.92\ 10^{-6}$	$2.7\ 10^{-8}$	$6.50\ 10^{-6}$	$4.5\ 10^{-8}$
S (m^2/g)	0.200	0.11	0.231	$9.7\ 10^{-3}$	0.177	$1.4\ 10^{-3}$	0.206	$2.6\ 10^{-3}$	0.346	$5.6\ 10^{-3}$	0.203	$3.2\ 10^{-3}$

instead of having $PZC = 6 \pm 1$ (Eq. 18) from experimental data.

On the other hand, another set found (Table 2c) is given with a small residual standard deviation, less than 0.1 for sorption of TBT and Na^+ constants, but still a large one of acidity constants (more than 0.3). In this case, we can assume that acidity constants are not yet well estimated, whereas sorption ones are quite accurate. Even if acidity constants are still not well estimated, the PZC is accurately estimated

$$PZC = \frac{1}{2}\left[\log(K_{\equiv S-OH_2^+}) - \log(K_{\equiv S-O^-})\right] = \frac{1}{2}(1.82 + 9.35) = 5.58 \quad (22)$$

Finally, the best set found from the *batch initial* set with $N_R = 2{,}000$ (Table 2d), is given with a small residual standard deviation, less than $4\ 10^{-2}$ for all equilibrium constants. The objective function is $F_N = 0.201$. The PZC of the sand is accurately given by the estimated parameters

$$PZC = \frac{1}{2}\left[\log(K_{\equiv S-OH_2^+}) - \log(K_{\equiv S-O^-})\right] = \frac{1}{2}(2.25 + 10.4) = 6.45 \quad (23)$$

We can see that the total sorption site estimated concentration, and the specific areas are close to the proposed one, and are given with a very small residual standard deviation.

By running a long research, that is $N_R = 20{,}000$, and $N_I = 1{,}000$, a better optimization is obtained. The objective function is small $F_N = 0.195$, and the proposed set (see Table 2f) is close to the best set of parameter given in Table 2e. The standard deviations associated to this set of parameters are small. Because the calculation of $N_R = 20{,}000$ realizations is very long (20 days), it is much more efficient to run two or three times $N_R = 2{,}000$ realizations, especially because the proposed sets are equivalent.

Uniqueness of the fit

In order to know if the best set found after $N_R = 2{,}000$ realizations (Table 2d) is the better one, a second optimization is run. The initial set of parameters is now the set given in Table 2d. The initial standard deviations are reduced, and are half of those given initially in Table 2a. After this second optimization, a better set is obtained, given in Table 2f. This set is the best one with an error function of $F_N = 0.192$. No more improvement is obtained by running a third optimization from the set of parameters given in Table 2e. All the parameters are given with a small standard deviation indicating that the convergence has been efficient. The PZC proposed for this set of parameter is $PZC = 5.24$. This value is quite far away from the experimental value of $PZC = 6 \pm 1$, but is still acceptable. A qualitative analysis of the parameters sensitivity can be done by comparing the different sets of parameters proposed in Table 2c, d, e, and f. The equilibrium constants for $\mathbf{K}_{\equiv S-OH_2^+}$, $\mathbf{K}_{\equiv S-OTBT}$, and $\mathbf{K}_{\equiv S-OHTBT^+}$, the total concentration $[\equiv S—OH]$, and the specific area S are always given with close values. On the other hand, equilibrium constants for $\mathbf{K}_{\equiv S-O^-}$ and $\mathbf{K}_{\equiv S-ONa}$ are very variable. These results are consistent with previous work.[21]

Table 3. Standard Deviation and Correlation Coefficients

	$\log(K_{=S-OH_2^+})$	$\log(K_{=S-O^-})$	$\log(K_{=S-OTBT})$	$\log(K_{=S-OHTBT^+})$	$\log(K_{=S-ONa})$	[≡S—OH] (mol/l)	S (m^2/g)
$\log(K_{=S-OH_2^+})$	**7,84E−03**						
$\log(K_{=S-O^-})$	−6,46E−02	**4.61E−03**					
$\log(K_{=S-OTBT})$	−8,85E−02	8,47E−01	**3,54E−03**				
$\log(K_{=S-OHTBT^+})$	5,34E−01	−8,02E−03	−5,81E−02	**5,11E−03**			
$\log(K_{=S-ONa})$	1,49E−01	−2,13E−02	1,29E−02	−4,16E−02	**5,78E−02**		
[≡S—OH] (mol/l)	−8,39E−02	−4,43E−01	−6,89E−01	−1,14E−01	1,54E−02	**6,08E−09**	
S (m^2/g)	−1,36E−01	−1,69E−01	−1,08E−01	−1,75E−02	−3,48E−02	1,61E−01	**1,64E−04**

Diagonal term (**bolt**) are the standard deviations for each parameter, nondiagonal one are correlations coefficients. We select 70 set of parameters with $F_N < 0.1923$.

This sensitivity analysis can be improved at low cost: from the optimization procedure, we get many set of parameters close to the optimal one. By selecting these which have an objective function sufficiently close to the best found (F_N = 0.1921), we obtain a representative set of accurate parameters. We select here the 70 set of parameters which have an objective function less than $F_N = 0.1923$. The standard deviation of these parameters and the correlation coefficients are given in Table 3. The standard deviation for each parameter is small. So we can assume that the best set of parameter is found with a sufficient precision. Because all the correlation coefficients are less than 0.9, the parameters are not correlated. This means that the hydrochemical system described in this work is not overparameterized. The standard deviation associated to $\mathbf{K}_{=S-ONa}$ in Table 3 is more or less 10 times larger than the standard deviation for other equilibrium constants. This confirms previous results obtained by Tovo[21]: the equilibrium constant $\mathbf{K}_{=S-ONa}$ does not have a significant influence on the objective function.

Breakthrough curves obtained using this best set of parameters is given in Figure 4. Comparing the results obtained after optimization (Figure 4) to these given from batch estimation (Figure 3) many improvement can be underlined. Elution curves at pH = 6.1 and 7.1 are much more similar according to experimental results. Retention of TBT at pH = 7.9 as been increased. A difference is now visible between elution at pH = 2.5, and elution at pH = 9.7 like for experimental curves. Injections curves have been changed a little bit. Injection at pH = 6.1 and 7.1 are much more similar, injection at pH = 5.2 is more delayed, and a difference is now visible between injection at pH = 2.5 and 9.7. Nevertheless, many problems are still present: even if injection curves seem to be slightly more diffusive, the calculated injection front is still compressive. Injection and leaching curves at pH = 7.9 are over delayed. Indeed, curves at pH = 7.9 are quite superposed with these obtained at pH = 6.1 and 7.1.

Conclusion

In this work, we develop a Monte-Carlo algorithm to optimize reactive transport parameters on multiconditional experiments. To use less computing time than a full random research, the mean and the standard deviation of the parameters are adapted during the optimization. This improvement allows us to use only 2,000 realizations of the reactive transport problem to find an accurate solution. Nevertheless, reducing computing time and the number of realizations introduces some hazard into the research procedure. It is also necessary to have some additional information, such as *PZC* value in our case, to validate the proposed set of parameters. In order to be sure of the proposed set of parameters, running many realizations is necessary too. As shown in Table 2, the results are not always

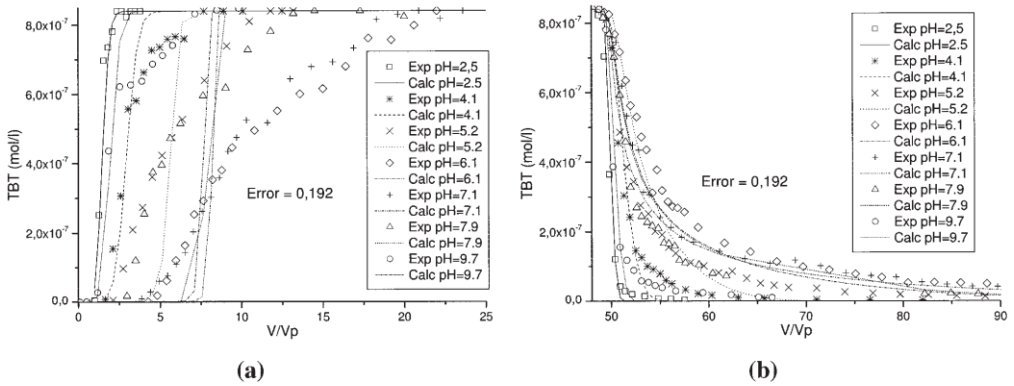


Figure 4. Experimental and optimized calculated elution curves.

Calculated elution curves are obtained after optimization with our algorithm. Estimated parameters obtained for N_R = 2,000 realizations as a second optimization from the best set found (Table 2d) at first optimization. F_N and $\mathbf{P}$ are given in Table 2e. Figure 4a: injection of TBT into the column. Figure 4b leaching of TBT out of the column.

the same. This is due to the reduction of the number of realizations, leading to a larger influence on the generated random numbers. Nevertheless, it is faster to run two or three times N_R = 2,000 realizations than to run one time N_R = 10,000 realizations, and leads to the same set of parameters. As shown by the correlation analysis, the set of parameters proposed by our algorithm is very accurate.

We have found many sets of parameters with an efficient convergence of the algorithm. Nevertheless, these sets of parameters are not the better one. This induces that the objective function has many local minima.

We have tried to find a set of parameters describing the seven injection-leaching experiments, assuming that the reactions governing the system are those given in Table 1. The best set of parameters found cannot describe accurately the experimental curves. This means that the proposed set of reactions does not represent the chemistry of TBT with the natural quartz sand. Bueno et al.[22] have shown, using experimental methods, that the sorption of TBT onto the natural quartz sand is better described by Langmuir-type isotherms with two sorption sites rather than by one. Nevertheless, a Langmuir isotherm can only describe the phenomena at one fixed pH, and does not take into account the electrostatic correction at the surface of the sand. The surface complexation model used in our work can describe the phenomena at various pHs, and takes into account the electrostatic correction. We can then answer the question posed previously, and we prove here that, even by describing more precisely the surface behavior during the sorption, one sorption site is not enough to represent accurately the sorption of TBT onto the natural quartz sand.

In this work, the natural quartz sand is described as an homogeneous surface with only one kind of sorption site: ≡S—OH. Analysis of this quartz sand[9,23] shows that the grain surface is composed of quartz, amorphous silica, ferric, and aluminum oxide and clay. According to our results, it seems that the macroscopic reactivity of a complex surface cannot be simplified to a unique sorption site if the pH is changing. This assumption has heavy consequences for the study of the sorption phenomena onto heterogeneous surface. Indeed, a macroscopic model, describing the sorption onto one site representing all the surface, will not be able to give an accurate overview of the phenomena, especially at various pHs. It is then necessary to produce more complex and complete sorption models, which take into account the heterogeneities of the surface. A further way of research will be to determine if all the heterogeneities have to be included into the sorption model, or if only the major ones are sufficient.

In order to increase the efficiency of the parameters estimation, more experimental information should be added. Only TBT elution curves have been used in this work. Sorption isotherms for batch experimentation can be added, or elution curves for other species. The effort done to reduce computing time must be continued, by increasing the efficiency of the chemical computation, and the reducing the number of time steps needed by transport model.

Acknowledgments

We thank Delphine Tovo (Ecole Nationale des Ingénieurs en Art Chimique et Technologique de Toulouse) for preliminary work during her engineer internship, and Ami Marxer for helpful comments. We thank Maïté Bueno for providing experimental data. M.A. has been supported by a grant from EGIDE.

Literature Cited

1. van der Lee J, De Windt L. Present state and future directions of modeling of geochemistry in hydrogeological systems. *J Contam Hydrol.* 2001;47:265–282.
2. Westall JC. FITEQL ver. 2.1. Corvallis, Department of Chemistry, Oregon State University; 1982.
3. Zhang T, Guay M. Adaptive parameter estimation for microbial growth kinetics. *AIChE J.* 2002;48:607–616.
4. Brassard P, Bodurtha P. A feasible set for chemical speciation problems. *Comps & Geosciences.* 2000;26:277–291.
5. Carrayrou J, Mosé R, Behra P. New efficient algorithm for solving thermodynamic chemistry. *AIChE J.* 2002;48:894–904.
6. Marshall SL. Generalized least-squares parameter estimation from multiequation implicit models. *AIChE J.* 2003;49:2577–2594.
7. Vikhansky A, Kraft M. A Monte Carlo methods for identification and sensitivity analysis of coagulation processes. *J Comput Phys.* 2004; 200:50–59.
8. Bueno M, Astruc A, Astruc M, Behra P. Dynamic sorptive behavior of tributyltin on quartz sand at low concentration levels: Effect of pH, flow rate, and monovalent cations. *Environ Sci Technol.* 1998;32: 3919–3925.
9. Bueno M. Etude dynamique des processus de sorption-désorption du tributylétain sur un milieu poreux d'origine naturelle. Université de Pau et des Pays de l'Adour, 1999. Ph.D. Thesis.
10. Carrayrou J. *Modélisation du transport de solutés réactifs en milieu poreux saturé.* Ph.D. thesis, Université Louis Pasteur Strasbourg I, 2001.
11. Sigg L, Behra P, Stumm W. *Chimie des milieux aquatiques.* Paris: DUNOD, 2000.
12. Morel FMM. *Principles of Aquatic Chemistry.* New York: Wiley Interscience, 1983.
13. Stumm W, Morgan JJ. Aquatic chemistry. Chemical equilibria and rates in natural waters. New York: Wiley-Interscience, 1996.
14. Dzombak DA, Morel FMM. Surface complexation modelling: Hydrous Ferric Oxide. New York; Wiley-Interscience, 1990.
15. Stumm W. Chemistry of the solid-water interface. New York: Wiley-Interscience, 1992.
16. Yeh GT, Tripathi VS. A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components. *Water Resour Res.* 1989;25:93–108.
17. Steefel CI, MQuarrie KTB. Approaches to modelling of reactive transport in porous media. In: Lichtner PC, Steefel CI, Oelkers EH, eds. *Reactive Transport in Porous Media.* Washington: Mineralogical Society of America, 1996:82–129.
18. Carrayrou J, Mosé R, Behra P. Modelling reactive transport in porous media: iterative scheme and combination of discontinuous and mixed-hybrid finite elements. *C. R. Acad Sci., Ser. II Univers.* 2003;331:211–216.
19. Hardyanto W. Groundwater modelling taking into account probabilistic uncertainties. Freiberg on-line Geosciences; 2003; 10.
20. Siegel P, Mosé R, Jaffré J. Solution of the advection dispersion equation using a combination of discontinuous and mixed finite elements. *Int J Numer Methods Fluids.* 1997;24:595–613.
21. Tovo D. Prise en compte de l'incertitude dans l'estimation des paramètres du transport réactif par une méthode Monte-Carlo. ENSIACET. Toulouse: France, Rapport de stage ENSIACET; 2003.
22. Bueno M, Astruc A, Lambert J, Astruc M, Behra P. Effect of solid surface composition on the migration of tributyltin in groundwater. *Environ Sci Technol.* 2001;35:1411–1419.
23. Behra P, Lecarme-Theobald E, Bueno M, Ehrhardt JJ. Sorption of tributyltin onto a natural quartz sand. *J Colloid Interface Sci.* 2003; 263:4–12.

Manuscript received Oct. 8, 2004, revision received Oct. 5, 2005, and final revision received Feb. 6, 2006.

Université de Strasbourg
Laboratoire d'**HY**drologie et de **GE**ochimie de **S**trasbourg
UMR 7517 CNRS UdS/EOST ENGEES
1, Rue Blessig 67 000 STRASBOURG, France
www.lhyges.unistra.fr

Résumé des travaux en vue d'obtenir
l'**Habilitation à Diriger des Recherches**

Modelling reactive transport in saturated porous media

Reactive transport modelling is a well-established research field. It is widely used for water resources management, contaminants flow prediction and comprehension, geological nuclear wastes disposal or CO_2 sequestration.
Reactive transport modelling requires a multiphysic description of the studied system, including the solid phase, the aqueous phase and the dissolved compounds; the transformation of each part and the interactions between all of them. Even if the models describing the water phase evolution and the aqueous chemical reactions are well known and accurate, the chemical reactions at the solid-water interface, their influence on the solid structure and on the water flow are still challenging because of the complexity of the phenomena, the heterogeneity of the porous media and the scaling effects. Transport phenomena are described through conservation equations leading to Darcy's law for water flow, advection-dispersion equation (that should be extended to Nernst-Planck one) for solute transport and advection-diffusion for heat transport. Chemical phenomena are described either by kinetic formulations for slow reactions or by instantaneous equilibrium for the fast ones. Instantaneous reactions at the liquid-solid interface can be mineral precipitation and dissolution, ion exchange, surface complexation with surface electrical potential or solid-solution formation.
We present some numerical works done to model reactive transport. In order to handle the multiphysical aspects of reactive transport, we chose an operator-splitting (OS) approach, which induces some errors. We present an analytical study of the OS schemes and the associated errors. Mass action laws and conservation equations are used to describe chemical equilibria. They lead to a non-linear system which is classically solved using the iterative Newton-Raphson method. Because of the very high non-linearity of the system, this method does not converge in some cases. We propose new numerical approaches to manage this problem, including numerical analysis of the system's condition number, matrix preconditioning and a new robust numerical method for non-linear systems: the Positive Continuous Fraction method. To model system including both equilibrium and kinetic reactions, we adapted the Richardson extrapolation to obtain a very flexible adaptive time step method.
Modern reactive transport codes are too complex to be verified but they can be tested and invalidated. By developing a set of reactive transport exercises for GdR MoMaS, we propose a benchmarking of seven reactive transport codes. This benchmark showed that the old classification (global approaches are less efficient that OS ones) is no more valid due to the new numerical methods and the evolution of the hardware.
We present two major orientations for future research: The first one aims at investigating the problematic of ion-specific diffusion and the respect of electro-neutrality and the second one deals with parameter estimation. When diffusion becomes the dominant transport process, the differences in transport between the diffusion coefficients of the ions have to be taken into account. We have to develop a model including electro-migration to maintain the electro-neutrality of the solution. The actual reactive transport codes are very complex and include a very large number of different reactive models. They require many reactive parameters that have to be fitted on experimental data. Nevertheless, because of the large number of parameters to be fitted, it is possible to describe an experiment using an erroneous reactive model. We expect to develop a global methodology and the required numerical tools for selecting the correct reactive model and estimating the parameters to describe an experiment or a set of experiments with a mechanistic reactive transport model.

Jérôme CARRAYROU

Jerome.carrayrou@unistra.fr

Recherche : Université de Strasbourg
Laboratoire d'HYdrogéologie et de GEochimie de Strasbourg
UMR CNRS UdS-EOST ENGEES 7517
1, rue Blessig 67 000 Strasbourg
Tél. : 03 90 24 29 16 Fax : 03 88 61 43 00
Poste direct 03 68 85 04 29

Enseignement : Université de Strasbourg
IUT Louis Pasteur de Schiltigheim
Département Génie biologique
1, allée d'Athènes 67300 Schiltigheim
Tél : 03 68 85 25 26 - Fax : 03 68 85 25 01
Poste direct 03 68 85 25 83