Sur la modélisation du transport réactif dans les réserves d'eaux potables

Safaa Al Nazer

Université Libanaise
Ecole Doctorale
Sciences et Technologie

Université
Littoral Côte d'Opale

THÈSE de doctorat

En vue de l'obtention du **Grade de Docteur**

délivré par

l'Université Libanaise - École Doctorale des Sciences et Technologie EDST

et

L'Université du Littoral Côte d'Opale - École Doctorale Science pour l'Ingénieur ED SPI

Sur la modélisation du transport réactif dans les réserves d'eaux potables

---

Spécialité : **Mathématiques et leurs interactions**

Réalisée par :
Safaa AL NAZER

Sous la direction de :
Pr. Mustapha JAZAR *Université Libanaise*
Pr. Carole ROSIER *ULCO*

Soutenue le 16 Décembre 2020, devant le jury :

Rapporteurs	Marwan FAHS	Université de Strasbourg
	Mladen JURAK	Université de Zagreb - Croatie
Président de jury	Mejdi AZAIEZ	Université Polytechnique de Bordeaux
Examinateurs	Jérôme CARRAYROU	Université de Strasbourg
	Hassan IBRAHIM	Université Libanaise
	Ghada CHMAYSEM	Université Libanaise
Directeurs	Mustapha JAZAR	Université Libanaise
	Carole ROSIER	ULCO

Cette thèse a été préparée aux laboratoires

LMPA Joseph Liouville
Maison de la Recherche Blaise Pascal
50, rue Ferdinand Buisson
CS 80699
62228 Calais Cedex
France
Téléphone : (33)(0)3 21 46 55 86,
Fax : (33)(0)3 21 46 55 75,
Email : secretariat@lmpa.univ-littoral.fr,
Site : www.lmpa.univ-littoral.fr/

LaMA-Liban
Université Libanaise
Centre Azm de Biotechnologie et ses Applications
Tripoli
Liban
Téléphone : (961) (03) 658 632,
Fax : (961) (06) 411 423,
P.O.Box 37 Tripoli via Beirut Lebanon,
Email : mjazar@ul.edu.lb,
Site : www.lamalb.net

Remerciements

Je tiens tout d'abord à exprimer ma profonde reconnaissance à Hassane Sadok, président de l'Université du Littoral Côte d'Opale et à Fawaz El Omar, doyen de l'école doctorale à l'Université Libanaise pour m'avoir offert une bourse doctorale qui m'a permis d'effectuer cette thèse en cotutelle. Cette expérience merveilleuse aura certainement une incidence sur le reste de ma vie.

Je tiens à remercier chaleureusement ma directrice de thèse en France Carole Rosier. Elle s'est toujours souciée de m'offrir, de tout point de vue, les meilleures conditions de travail possibles. Son encadrement scientifique m'a permis de préparer cette thèse dans d'excellentes conditions. Elle m'a apporté les outils nécessaires pour devenir une chercheuse indépendante grâce à son écoute et à sa précieuse aide. Qu'elle trouve ici l'expression de ma profonde reconnaissance.

Je voudrais également remercier mon directeur de thèse au Liban Mustapha Jazar. Merci pour son aide exceptionnelle sur les plans humain, scientifique et administratif. Merci pour ses qualités d'écoute, sa présence à mes côté et ses conseils avisés et rassurants. J'espère avoir été digne de la confiance qu'il m'a accordée et que ce travail est finalement à la hauteur de ses espérances. Je suis très honorée de l'avoir eu pour encadrant.

Je remercie tout particulièrement Marwan Fahs et Mladen Jurak pour avoir accepté d'être rapporteurs de ma thèse ainsi que d'être membres du jury. Je les remercie pour le soin qu'ils ont apporté à la lecture de ce document ainsi que pour les commentaires constructifs et avisés qui ont permis d'améliorer mon manuscrit de thèse.

Je tiens également à remercier Jérôme Carrayrou pour tous les échanges que nous avons eus, notamment lors de la réalisation de la première partie de mon travail. Je le remercie aussi de bien avoir voulu faire partie du jury.

J'exprime ma profonde gratitude à Mejdi Azaïez pour m'avoir fait l'honneur de présider mon jury de thèse et s'être intéressé à mon travail.

Je remercie sincèrement Hassan Ibrahim et Ghada Chmaysem pour l'intêret qu'ils ont porté à ma thèse et pour avoir accepté de faire partie de mon jury de thèse.

Ma plus profonde gratitude va à mon mari Amjad El Yassin, à mon père Mahmoud Al Nazer et ma mère Zaynab Al Nazer, à mes sœurs et mes frères pour leur soutien continu et pour m'avoir donné l'énergie de persévérer dans les échecs et de surmonter avec succès les nombreuses difficultés rencontrées.

Enfin je ne peux terminer sans parler des beaux évènements que j'ai vécus ces trois années : je suis devenue mère de deux petits anges "Al Houmam" et "Hala". Ils m'ont donné toujours l'espoir de réussir et de réaliser un meilleur travail. J'ai rencontré beaucoup de difficultés pendant cette période, mais à chaque fois, je puisais la force dans leurs yeux afin de surmonter l'échec et ainsi faire de mon succès un bel avenir pour eux.

Résumé

Cette thèse est consacrée à l'étude du transport réactif dans les réserves en eaux. Elle est structurée en deux parties distinctes : la première porte sur l'élaboration de solveurs chimiques et la seconde sur l'étude mathématique d'une classe de modèles décrivant des écoulements en eaux peu profondes en interaction avec les eaux de surface.

Dans la première partie du travail, on s'intéresse à la résolution numérique des équilibres thermodynamiques qui conduisent à des systèmes non linéaires complexes et très mal conditionnés. Dans ce travail, on combine une formulation particulière du système d'équilibre chimique, appelée la méthode des fractions continues positives, avec deux méthodes numériques itératives, la méthode d'Accélération d'Anderson et des méthodes d'extrapolation vectorielle, à savoir les méthodes MPE (minimal polynomial extrapolation) et RRE (reduced rank extrapolation). Le principal avantage de ces approches est d'éviter de former la matrice jacobienne et donc d'éviter les problèmes liés aux mauvais conditionnements de la matrice. Des tests numériques sont faits, notamment sur le cas test de l'acide gallique et sur le cas test $1D$ de référence du benchmark MoMas. Ces essais illustrent la grande efficacité de cette approche par rapport aux résolutions classiques résultant de la méthode de Newton-Raphson.

Dans la seconde partie de la thèse, on introduit et étudie des modèles de type Richards-Dupuit pour décrire les écoulements dans des aquifères peu profonds. L'idée est de coupler les deux types d'écoulements principaux présents dans l'aquifère : celui de la partie insaturée avec celui de la partie saturée. Le premier est décrit par le problème classique de Richards dans la frange capillaire supérieure. Le second résulte de l'approximation de Dupuit après intégration verticale des lois de conservation entre le fond de l'aquifère et l'interface de saturation. Le modèle final consiste en un système fortement couplé d'edp de type parabolique qui sont définies sur un domaine dépendant du temps.

Nous montrons comment la prise en compte de la faible compressibilité du fluide permet d'éliminer la dégénérescence présente dans la dérivée temporelle de l'équation de Richards. Puis nous utilisons le cadre général des équations paraboliques dans des domaines non cylindriques introduit par Lions pour donner un résultat d'existence global en temps. Nous présentons l'analyse mathématique du premier modèle qui correspond au cas isotrope et non conservatif. Puis nous généralisons l'étude au cas anisotrope et conservatif.

Abstract

This thesis is devoted to the study of reactive transport in water reserves. It is structured in two distinct parts : the first deals with the development of chemical solvers and the second with the mathematical study of a class of models describing flows in shallow water interacting with the surface water.

In the first part of the work, we focus on the numerical resolution of thermodynamic equilibria which lead to complex and very badly conditioned nonlinear systems. In this work, we combine a particular formulation of the chemical equilibrium system, called the method of positive continuous fractions, with two iterative numerical methods, the Anderson Acceleration method and vector extrapolation methods, namely the MPE (minimal polynomial extrapolation) and RRE (reduced rank extrapolation) methods . The main advantage of these approaches is to avoid forming the Jacobian matrix and thus to avoid problems linked to bad conditioning of the matrix. Numerical tests are performed, especially on the test case of gallic acid and on the reference $1D$ case of the MoMas benchmark. These tests illustrate the great efficiency of this approach compared to classical solutions resulting from the Newton-Raphson method.

In the second part of the thesis, we introduce and study Richards-Dupuit type models to describe flows in shallow aquifers. The idea is to couple the two main types of flows in the aquifer : that of the unsaturated part with that of the saturated part. The first is described by the classic Richards problem in the upper capillary fringe. The second results from Dupuit's approximation after vertical integration of the conservation laws between the bottom of the aquifer and the saturation interface. The final model consists of a strongly coupled system of parabolic type pde which are defined on a time dependent domain. We show how taking into account the low compressibility of the fluid makes it possible to eliminate the degeneration in the time derivative term of the Richards equation. Then we use the general framework of parabolic equations in non-cylindrical domains introduced by Lions to give a global existence result in time. We present the mathematical analysis of the first model which corresponds to the isotropic and non-conservative case. Then we generalize the study to the anisotropic and conservative case.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte général

Dans le cycle de l'eau, les eaux souterraines jouent un rôle primordial et constituent une réserve en eau potable non négligeable. L'évolution des polluants qui peuvent s'infiltrer jusqu'à la surface des nappes d'eau souterraines (citons parmi ceux-ci les produits radioactifs, déchets toxiques ou autres contaminants, issus des activités humaines, industrielles ou agricoles) est très lente, comparativement à ce qui se passe en surface. Il s'écoule des mois, des années, ou même des dizaines d'années, entre le début de la pollution et sa mise en évidence. Les nuisances peuvent ainsi se maintenir très longtemps après le tarissement des sources de pollution. L'éventualité d'un impact des activités humaines sur l'environnement et la qualité des eaux a suscité un intérêt croissant du monde scientifique et explique pourquoi le *transport réactif* est devenu un domaine de recherche important au cours des dernières décennies. En effet, la mise en place de sites d'entreposage de déchets sûrs, le développement de méthode de dépollution des nappes polluées, l'amélioration des stratégies de prévention de la pollution nécessitent une meilleure description et prédiction du devenir des contaminants dans les milieux poreux souterrains. C'est ainsi que la *modélisation* apparaît comme un outil intéressant permettant d'évaluer nos connaissances en essayant de reproduire les phénomènes observés et de prévoir le devenir des contaminants dans les sols afin de mieux comprendre les processus mis en jeu. C'est dans cette optique que nous apportons, par ce travail, une contribution à une stratégie efficace utilisée pour une meilleure modélisation du transport et de la chimie. Pour ce faire, le modèle de transport doit prendre en compte les processus physiques à travers les équations de conservations de la masse (équations aux dérivées partielles) et le modèle de la chimie prend en compte les processus chimiques ou biologiques à travers les lois d'action de masse. Dans ce travail, les deux aspects "chimie" et "écoulement engendrant le transport" de la problématique du transport réactif sont décrits et étudiés séparément. Le contexte chimique constitue la première partie du document tandis que le contexte lié au transport constitue la deuxième partie.

Au cours des dernières décennies, le transport réactif était considéré comme un sujet majeur dans de nombreux domaines scientifiques tels que la combustion, la catalyse, la mécanique des fluides, le génie chimique et la géochimie. Les flux réactifs à plusieurs composants monophasés sont modélisés par une loi d'équilibre de masse, la loi de Darcy et des équations d'état. Dans le cas des réactions d'équilibre, les lois d'action de masse consistent en des équations algébriques reliant les activités des espèces impliquées. Le problème de transport réactif est ainsi modélisé par des équations aux dérivées partielles décrivant l'écoulement couplées à des équations algébriques décrivant des réactions chimiques. En raison de la complexité des systèmes et de la non-linéarité des processus chimiques, le transport réactif à plusieurs composants entraîne une exigence de calcul importante. Dans

ce contexte, deux stratégies numériques sont généralement utilisées pour résoudre ce système : l'approche implicite global (GIA) et l'approche séquentiel itératif (et non itératif) (SIA), également appelé approche de fractionnement d'opérateurs (voir par exemple les références [25, 80, 106, 107]). L'approche GIA résout à chaque pas de temps le système non linéaire complet résultant de la substitution directe des équations chimiques dans les équations de transport tandis que l'approche SIA résout séquentiellement les équations de transport et les réactions biogéochimiques. Les résultats de comparaisons récentes entre GIA et SIA obtenus par différentes équipes sont en bon accord [39], comme ceux donnés en [108] pour lesquels une méthode de volumes finis totalement implicite a été développée et mise en œuvre dans le cadre de la plate-forme open-source parallèle Dumu^X ([15, 110]). Ces différents benchmarks ont montré que la précision des approches séquentielles est comparable à celle des approches globales et que ces dernières sont désormais plus efficaces qu'on ne le croyait à l'origine (même si dans certains travaux, il est mentionné que l'approche GIA est beaucoup plus coûteuse en temps de calcul et stockage que l'approche SIA (cf. [82])). Chacune de ces méthodes présente des qualités et des inconvénients mais quelle que soit l'approche, un problème non linéaire doit être résolu par une méthode de point fixe et la méthode de Newton Raphson est souvent utilisée pour cette résolution numérique. Cependant, la résolution de tels systèmes non linéaires, notamment due à des processus chimiques, peut conduire à une non convergence ou à un nombre excessif d'itérations en raison de la nature très mal conditionnée du problème. Le but de ce travail est de proposer de nouveaux algorithmes puissants (en termes de temps CPU et de stabilité) qui permettront de faire face à ces problèmes difficiles.

La chimie des milieux aquatiques naturels se caractérise par une grande diversité, tant en ce qui concerne les phénomènes chimiques que les espèces chimiques, les ordres de grandeurs ou les conditions réactionnelles. De nombreux critères de classement ont été proposés pour ordonner les multiples phénomènes chimiques, citons la vitesse de la réaction (réaction à l'équilibre thermodynamique instantané ou contrôlée par les lois cinétiques) et le nombre de phases mises en jeu lors de la réaction (réaction homogène ou hétérogène, phase mobile ou immobile). L'hypothèse qu'une réaction est à l'équilibre thermodynamique constitue une base de la modélisation chimique adoptée dans le cadre de ce travail. Cette hypothèse permet de bénéficier de bases de données très riches et de réduire le nombre de variables primaires du système. De plus, toutes les phases peuvent être traitées de façon similaire (bien qu'adaptées à leurs propriétés individuelles). En revanche, il est impératif et évident de distinguer, d'un point de vue hydrodynamique, les phases mobiles des phases immobiles et par suite les espèces mobiles des espèces immobiles. D'autre part, il est préférable d'utiliser une représentation des systèmes chimiques en faisant une distinction entre les espèces chimiques primaires et les espèces chimiques secondaires. Pour ce faire, nous utilisons, comme la plupart des chimistes [15, 17, 20], le tableau des équilibres de Morel [27] pour une représentation claire de l'ensemble des réactions du système chimique.

En termes thermodynamiques, un calcul de l'équilibre chimique, qui tente de trouver la valeur minimale de l'énergie libre de Gibbs, peut être effectué de deux manières : en minimisant une fonction d'énergie libre ou en résolvant un ensemble d'équations non linéaires constituées de constantes d'équilibre et des contraintes de bilan massique. Notons que, dans le contexte de l'industrie pétrolière, une approche alternative récente [103] étudie l'équilibre de phase sous un volume fixe plutôt qu'une pression fixe et minimise l'énergie libre de Helmholtz au lieu de l'énergie libre de Gibbs. Enfin, des travaux récents utilisent des algorithmes d'apprentissage profond efficaces pour estimer les états d'équilibre thermodynamique de fluides de réservoir réalistes avec un grand nombre de composants permettant ainsi d'accélérer les calculs d'équilibre de phase. Plus précisément, une stratégie d'accélération simple réduit le nombre de composants dans le mélange fluide améliorant l'efficacité des algorithmes sans compromettre la précision des équations d'états (voir [104, 105]). Ces méthodes sont thermodynamiquement équivalentes, mais le principal inconvénient de l'utilisation d'une base de données d'énergie libre est que ces valeurs ne sont pas aussi fiables que les constantes d'équilibre mesurées directement. La précision des résultats des solveurs chimiques étant particulièrement re-

cherchée, surtout si l'on veut les intégrer dans les méthodes SIA, nous allons nous concentrer sur la résolution numérique des équations non linéaires décrivant les équilibres thermodynamiques.

Devant l'impossibilité de mettre en œuvre une résolution directe, dans le cas général, du système algébrique non linéaire définissant l'équilibre thermodynamique du système chimique, il est nécessaire de faire appel à des méthodes itératives. De nombreuses méthodes mathématiques ont été testées pour résoudre un tel système algébrique non linéaire. Comme déjà mentionné ci-dessus, la méthode de Newton-Raphson est la plus utilisée pour calculer l'équilibre thermodynamique ou plus généralement pour résoudre l'ensemble des équations non linéaires. avec la difficulté que la matrice jacobienne engendrée par cette approche doit être calculée, stockée, factorisée et est généralement très mal conditionnée, ce qui nécessite des procédures de préconditionnement [37, 38]. Cela peut devenir problématique pour les grands problèmes. De plus, la localisation des données initiales dans tout algorithme de type Newton est une difficulté récurrente qui ralentit et même empêche la convergence de l'algorithme. Enfin, même des systèmes chimiques petits ou très petits (4×4 à 20×20, parfois plus grands) peuvent être très mal conditionnés (nombre de conditionnement jusqu'à 10^{100}) (cf. [38]).

Afin d'étudier cette problématique, nous cherchons dans ce travail à résoudre le système chimique à l'équilibre thermodynamique par d'autres méthodes ne nécessitant pas le calcul de la matrice jacobienne. Notre approche consiste tout d'abord à transformer le problème chimique en un problème de point fixe. En effet, il existe une dualité naturelle entre un problème de point fixe et un problème d'équations non linéaires. Nous nous sommes particulièrement concentrés sur trois méthodes d'accélération itérative : la méthode d'*Accélération d'Anderson* AA [1, 32] et les deux méthodes d'extrapolation vectorielle de type polynomial MPE (*Minimal Polynomial Extrapolation*) de Cabay et Jackson [12] et RRE (*Reduced Rank Extrapolation*) d'Eddy [13] et MeSina [14]. Ces méthodes n'ont jamais été appliquées à la résolution des équilibres thermodynamiques. De plus, leur efficacité est améliorée en les associant à une formulation particulière du système d'équilibre : la méthode des *Fractions Continues Positives* (PCF) introduite par Jérome Carrayrou [22]. Habituellement, la méthode (FCP) est utilisée pour préconditionner la méthode de Newton Raphson pour les espèces majeures (comme dans le PHREEQC [16]) ou pour réduire les difficultés dues au manque de convergence globale de la méthode de Newton, si la condition initiale n'est pas suffisamment proche de la solution (voir [22]). La combinaison directe de la méthode (PCF) avec AA, RRE ou MPE présentée dans ce travail fournit des algorithmes très efficaces et robustes avec une convergence super linéaire ou quadratique à partir de toute donnée initiale arbitraire. Des tests numériques sont faits notamment sur le cas test de l'acide gallique et sur le cas test "$1D$ Easy"" du Benchmark MoMas. Ces expériences numériques justifient l'efficacité de l'approche suivie ainsi que la robustesse des algorithmes des méthodes itératives utilisées. Ils illustrent la grande efficacité de notre approche en faisant une comparaison systématique avec des résolutions résultant de la méthode classique de Newton-Raphson.

Le couplage entre les opérateurs de transport et ceux de chimie est un des principaux objectifs concernant les travaux ultérieurs liés à ce travail. La méthode développée dans le cadre de cette thèse peut être appliquée dans des codes basés sur l'approche séquentielle. Ce point constitue en grange partie une des motivations majeures de ce travail.

Dans la seconde partie de la thèse, on s'intéresse à l'aspect "écoulement engendrant le transport" de la problématique du transport réactif. On se concentre sur l'aspect hydrogéologique des aquifères. Les sources de pollution sont à l'origine de la contamination des sols et de la détérioration des aquifères d'eau douce. Ainsi, l'étude des écoulements d'eau dans les milieux poreux (saturés et non saturés) devient un enjeu important pour la consommation d'eau dans plusieurs domaines entre autre tels que l'agriculture, l'environnement, la gestion des déchets, le développement urbain et les processus industriels.

Par ailleurs, la description quantitative du processus d'écoulement dans les aquifères souterrains devient très complexe, puisqu'allant d'une partie saturée à une partie non saturée, du fait de leur spécifité et à cause des variations de l'état hydrique du sol pendant l'écoulement. Ces variations im-

pliquent des relations complexes entre les différents paramètres de l'écoulement. Ceci oblige la plupart des scientifiques et des ingénieurs à s'appuyer fortement sur des méthodes d'analyse mathématique basées sur des approches expérimentales afin d'obtenir des modélisations des écoulements souterrains permettant de mieux appréhender le comportement du débit d'eau dans les aquifères. La modélisation mathématique des écoulements souterrains a ainsi connu un grand essor au cours des dernières décennies. Les modèles issus sont basés sur des hypothèses simplificatrices constituant une approximation des conditions sur le terrain. Toutefois, même si elles sont approchées, ces hypothèses fournissent un outil d'investigation fondamental que les hydrogéologues peuvent utiliser dans diverses applications, telles que la migration chimique dans les zones saturées et les échanges entre les rivières et les eaux souterraines.

La loi de Darcy formulée par Henry Darcy pour la première fois en 1856 reste de nos jours un élément essentiel de la description mathématique de l'écoulement d'un fluide dans un milieu poreux. Elle permet d'exprimer de manière remarquablement simple la vitesse de l'écoulement en fonction du gradient de pression. Elle est toujours largement utilisée dans de nombreux domaines (hydrologie, génie chimique, exploitation des gisements d'hydrocarbures...). Lorenzo Adolph Richards a ensuite contribué au second apport décisif, au-delà des travaux d'Edgar Buckingham, en généralisant la loi de Darcy aux écoulements dans des sols non saturés. Cela a donné lieu à une équation aux dérivées partielles, connue aujourd'hui sous le nom d'*équation de Richards* décrivant l'évolution de la teneur en eau du sous-sol.

Notre travail traite des phénomènes naturels liés à ces questions hydrogéologiques. En général, le mouvement des eaux souterraines est considéré comme un problème de fluides polyphasiques qui est donc décrit par les équations de Richards [69]. Il s'agit d'un système tridimensionnel d'équations aux dérivées partielles non linéaires dégénérées de type parabolique. Notre travail repose principalement sur ces équations. Nous nous concentrons sur l'écoulement d'un fluide faiblement compressible (l'eau) dans des aquifères, grands et peu profonds. Nous observons qu'il y a deux types d'écoulements dominants présents dans l'aquifère : celui de la partie insaturée et celui de la partie saturée. L'idée est de coupler ces deux types d'écoulements. Le premier apparaît dans la frange capillaire supérieure et est décrit par le problème classique de Richards. Le second apparaît dans la région inférieure de l'aquifère et se fait globalement dans la direction horizontale. Ces deux composantes de l'écoulement sont séparées par une interface qui est donc l'intersection de ces deux régions. La région inférieure est complètement saturée tandis que la partie supérieure est partiellement saturée (et pouvant être parfois sèche). En particulier au-dessus de l'interface, le flux vertical est dominant alors qu'en dessous, il est quasiment instantané. Ainsi, l'hypothèse de Dupuit est satisfaite dans la région inférieure du réservoir. Cela permet l'intégration verticale des lois de conservation dans cette partie et conduit à l'utilisation d'une famille de modèles $2D$ fortement développés depuis les années 60 (voir, par exemple, les travaux de Jacob Bear, [42, 43]). Le modèle final consiste donc en un système fortement couplé d'équations aux dérivées partielles de type parabolique qui sont définies sur un domaine dépendant du temps (notamment de l'interface entre les deux sous-domaines). Dans [48], une classe de tels modèles est proposée. Elle consiste à coupler des modèles purement verticaux décrivant l'écoulement en temps courts dans la frange capillaire, avec un modèle horizontal décrivant l'écoulement en temps longs dans la zone saturée. Ce modèle est une alternative au problème de Richards $3D$ pour décrire l'écoulement dans un aquifère peu profond pour une large gamme d'échelles de temps. Dans ce travail, notre étude se base sur cette classe de modèles car elle présente deux avantages par rapport au modèle de Richards $3D$ dont elle est issue. Plus précisément, elle génère des codes numériques plus rapides à résoudre offrant ainsi un gain de temps CPU important, tout en reproduisant les mêmes comportements asymptotiques dominants que le problème original de Richards $3D$ lorsque le rapport profondeur/largeur de l'aquifère est petit, et ceci pour toute échelle de temps considérée. Par contre, l'analyse mathématique d'un tel modèle s'avère très délicate et rencontre des difficultés liées à l'intégration des équations sur un domaine à frontière libre, aux nonlinéarité et dégénérescences dans les dérivées en temps et en espace des équations et à la perte de contrôle des composantes ho-

rizontales du gradient de pression.
Usuellement, la transformée de Kirchoff, appliquée à l'équation de Richards, permet d'éliminer la nonlinéarité dans le terme diffusif. Dans ce travail, nous montrons comment la prise en compte de la faible compressibilité du fluide permet à son tour d'éliminer la dégénérescence présente dans la dérivée temporelle de l'équation de Richards. Puis nous utilisons le cadre général des équations paraboliques dans des domaines non cylindriques introduit par Mignot et Lions [57, 58] pour traiter le problème à frontière libre et donner un résultat d'existence global en temps.
Deux analyses mathématiques sont présentées dans ce travail dépendant des modèles considérés : le premier correspond au cas isotrope et non conservatif pour lequel la conductivité hydraulique est moyennée verticalement. Le second correspond au cas anisotrope et conservatif pour lequel, comme dans [48], nous considérons une forme très générale pour la conductivité hydraulique. Ce dernier modèle est construit de sorte qu'il soit très proche, du point de vue physique, du modéle donné dans [48] tout en introduisant une faible conductivité hydraulique horizontale dans la partie insaturée de l'aquifère, permettant ainsi de pallier à la perte d'information sur les composantes horizontales du gradient de pression.

Les résultats de ce travail sont exposés sous forme de trois articles (correspondant aux trois chapitres suivants), ils sont rédigés en anglais. L'introduction rédigée en français donne un aperçu détaillé de ces résultats.

1.2 Modélisation d'un système chimique pour une réaction à l'équilibre thermodynamique

L'objectif d'un module de spéciation chimique est de calculer les concentrations des différentes espèces présentes dans le système considéré, à partir de la connaissance des réactions, et des éléments chimiques en cause.
Les réactions chimiques peuvent être classifiées selon deux critères : en fonction du caractère homogène ou non de la réaction, ou en fonction de la vitesse de la réaction.

Suivant la vitesse, la réaction peut être considérée à l'équilibre thermodynamique instantané ou contrôlée par la loi cinétique chimique. Quand la vitesse de réaction est lente par rapport à celle du transport, ou du même ordre que celle-ci, on doit prendre en compte la cinétique de la réaction. Dans le cas contraire, le système est réversible, on peut considérer qu'il est à l'*équilibre à chaque instant*, et qu'une modélisation par des réactions à l'équilibre est adéquate.

Le nombre de phases mises en jeu par une réaction chimique n'est pas un élément déterminant d'un point de vue strictement chimique. Les *réactions hétérogènes* font intervenir deux ou plusieurs phases (adsorption, précipitation-dissolution) alors que les *réactions homogènes* (oxydo-réductions, acido-basiques, complexation) se produisent dans une même phase.
Plus généralement, les réactions homogènes correspondent aux modèles d'écoulements souterrains avec des réactions faites en phase aqueuse (l'eau). Les réactions hétérogènes jouent évidemment un rôle fondamental en géochimie. Elles comprennent des réactions surfaciques et de précipitation [21].
Les réactions surfaciques sont les réactions où les espèces chimiques dissoutes interagissent avec la surface de la matière rocheuse par des processus de sorption, par lesquelles une espèce aqueuse peut former un solide complexe. Les réactions de précipitation représentent les phénomènes de dissolution des minéraux dans l'eau, qui peuvent exister sous forme solide ou dissoute.
Les espèces chimiques sont reliées entre eux par les différentes réactions chimiques possibles. Ensuite, on peut exprimer certaines espèces en fonction des autres, ce qui permet de réduire la taille du

système à résoudre. Dans la suite, nous nous restreignons aux réactions équilibrées. On parle alors d'espèces primaires et secondaires pour formuler un système chimique à l'équilibre.

1.2.1 Présentation

La description des réactions chimiques en équilibre thermodynamique peut être présentée par l'étude d'une réaction simple (1.1) entre deux espèces $\mathcal{E}_1$ et $\mathcal{E}_2$:

$$\mathcal{E}_1 \rightleftarrows \mathcal{E}_2. \tag{1.1}$$

À l'équilibre, le système chimique est décrit par les relations suivantes :

- L'équation de conservation de la matière impose l'égalité entre la somme des concentrations $[\mathcal{E}_1]$ et $[\mathcal{E}_2]$ des deux espèces et la quantité totale de réactif injectée dans le système $[T]$:

$$T = [\mathcal{E}_1] + [\mathcal{E}_2]. \tag{1.2}$$

- La loi d'action de masse, obtenue en minimisant l'enthalpie libre de la réaction (1.1) ΔG° ou Energie de Gibbs du système [23] donne la relation entre les activités $\{\mathcal{E}_1\}$ et $\{\mathcal{E}_2\}$ à l'équilibre thermodynamique des deux espèces $\mathcal{E}_1$ et $\mathcal{E}_2$:

$$K = \frac{\{\mathcal{E}_1\}}{\{\mathcal{E}_2\}}, \tag{1.3}$$

où K est la constante d'équilibre thermodynamique. K dépend de ΔG°, à la pression P et à la température T° :

$$K = -\frac{\Delta G^\circ(P, T^\circ)}{RT^\circ}. \tag{1.4}$$

Cette approche permet une prise en compte directe des variations de pression et de température sur le système [24]. Mais, d'autre part, cette approche souffre d'un manque de données directement exploitables, les bases de données géochimiques [17] étant préférentiellement présentés pour la loi d'action de masse (1.3). Pour ces raisons, notre travail ici a été développé selon l'approche habituelle des modèles géochimiques où le calcul des équilibres thermodynamiques se fait par la loi d'action de masse.
Comme la plupart des modélisateurs [15, 17, 20], on préfère utiliser la représentation des sytèmes chimiques par espèces primaires et secondaires. L'utilisation du tableau des équilibres de Morel permet alors une écriture claire de l'ensemble des réactions du système chimique.

1.2.2 Espèces primaires et espèces secondaires

On considère un ensemble de n_r réactions chimiques entre n_e espèces chimiques $\mathcal{E}_j$, $j = 1, ..., n_e$, tel que $n_r \leq n_e$

$$\sum_{j=1}^{n_e} \tilde{\mu}_{ij} \mathcal{E}_j \rightleftarrows 0 \qquad i = 1, ..., n_r, \tag{1.5}$$

où $\tilde{\mu}_{ij}$ est la matrice stoechiométrique des espèces $\mathcal{E}_j$ dans la réaction i. (1.5) s'écrit sous la forme matricielle suivante

$$\tilde{\mu}\mathcal{E} \rightleftarrows 0. \tag{1.6}$$

Après substitutions et reformulations, chaque réaction peut être écrite sous une forme donnant un seul produit distinct par réaction. Donc, la matrice stoechiométrique $\tilde{\mu}$, étant de rang maximal n_r, peut s'écrire sous la forme échelon $\tilde{\mu} = [-I_{n_r} \quad \mu]$, où I_{n_r} est la matrice identité de taille n_r. On peut alors écrire le système chimique sous la forme

$$\mathcal{C}_i \rightleftarrows \sum_{j=1}^{n_e - n_r} \mu_{ij} \mathcal{X}_j \qquad i = 1, ..., n_r \tag{1.7}$$

ou encore sous forme matricielle

$$\mathcal{C} \rightleftarrows \mu^T \mathcal{X}. \tag{1.8}$$

Les espèces dénotées à gauche comme produits, $\mathcal{C}$, sont dites des ***espèces secondaires*** et les espèces à droite, $\mathcal{X}$, sont dites ***espèces primaires*** ou ***composants***. Ainsi, l'équation (1.7) montre que la formation des espèces secondaires $\mathcal{C}$ a lieu à partir des espèces primaires $\mathcal{X}$, d'une façon unique. L'avantage de cette approche est qu'elle réduit la taille du système chimique à résoudre. On caractérise encore les espèces suivant le type de phase auxquelles elles appartiennent : une espèce est dite ***mobile*** (m) si elle appartient à une phase mobile, ***fixée*** ou ***immobile*** (f) si elle appartient à une phase immobile et précipitée (π) si elle est minérale. En utilisant les notations :

- X : sous ensemble des espèces primaires mobiles de cardinal n_{pm},
- S : sous ensemble des espèces primaires fixées de cardinal n_{pf},
- C : sous ensemble des espèces secondaires mobiles de cardinal n_{sm},
- CS : sous ensemble des espèces secondaires fixées de cardinal n_{sf},
- π : sous ensemble des espèces précipitées de cardinal n_π,
- $\mu^{(C,X)} \in \mathbb{R}^{n_{sm} \times n_{pm}}$: bloc de la matrice stœchiométrique entre C et X,
- $\mu^{(\pi,X)} \in \mathbb{R}^{n_\pi \times n_{pm}}$: bloc de la matrice stœchiométrique entre π et X,
- $\mu^{(CS,X)} \in \mathbb{R}^{n_{sf} \times n_{pm}}$: bloc de la matrice stœchiométrique entre CS et X,
- $\mu^{(CS,S)} \in \mathbb{R}^{n_{sf} \times n_{pf}}$: bloc de la matrice stœchiométrique entre CS et S,

on donne une représentation synthétique du système chimique de la manière suivante

$$\begin{pmatrix} & \mu^{(C,X)} & 0 \\ -I_{n_r} & \mu^{(CS,X)} & \mu^{(CS,S)} \\ & \mu^{(\pi,X)} & 0 \end{pmatrix} \begin{pmatrix} C \\ CS \\ \pi \\ X \\ S \end{pmatrix} \rightleftarrows 0 \tag{1.9}$$

avec

$$\mu = \begin{pmatrix} \mu^{(C,X)} & 0 \\ \mu^{(CS,X)} & \mu^{(CS,S)} \\ \mu^{(\pi,X)} & 0 \end{pmatrix}, \quad \mathcal{C} = \begin{pmatrix} C \\ CS \\ \pi \end{pmatrix}, \quad \mathcal{X} = \begin{pmatrix} X \\ S \end{pmatrix} \quad \text{et} \quad \mathcal{E} = \begin{pmatrix} \mathcal{X} \\ \mathcal{C} \end{pmatrix}.$$

Notons que les espèces primaires fixées ne prennent pas part aux réactions homogènes (définies ci après) qui impliquent seulement les espèces mobiles et que les réactions de précipitation n'impliquent pas les espèces fixées. Dans ce travail, on s'intéresse aux systèmes chimiques sans espèces précipitées, c'est à dire que $\pi = \phi$ et $\mu^{(\pi,X)} = 0$. Un algorithme classique [22] pour décrire la précipitation ou la dissolution des minéraux fait une hypothèse a priori sur l'existence ou la non-existence de minéraux. Dans le cadre de ce travail, on suppose que cette hypothèse est proposée.

Dans ce qui suit, on désigne par $X_1, ..., X_{n_{pm}}$ les espèces primaires mobiles et par $S_1, ..., S_{n_{pf}}$ les espèces primaires fixées. De même, on désigne par $C_1, ..., C_{n_{sm}}$ les espèces secondaires mobiles et par $CS_1, ..., CS_{n_{sf}}$ les espèces secondaires fixées. Ainsi, en notant par $\mu_1 = \mu^{(C,X)}$, $\mu_2 = \mu^{(CS,X)}$ et $\mu_3 = \mu^{(CS,S)}$, on distingue les réactions chimiques comme suit :

i- Réactions entre espèces mobiles

$$\sum_{j=1}^{n_{pm}} \mu_{1,i,j} X_j \rightleftarrows C_i \qquad i = 1, ..., n_{sm} \tag{1.10}$$

ii- Réactions entre espèces mobiles et fixées (réactions de sorption)

$$\sum_{j=1}^{n_{pm}} \mu_{2,i,j} X_j + \sum_{j=1}^{n_{pf}} \mu_{3,i,j} S_j \rightleftarrows CS_i \qquad i = 1, ..., n_{sf} \tag{1.11}$$

1.2.3 Loi d'action de masse

L'introduction de la loi d'action de masse dans le système chimique conduit à un système plus complexe et difficile à résoudre. Étant données les concentrations des espèces primaires, cette loi décrit comment on peut obtenir les concentrations des espèces secondaires. Le formalisme de la loi d'action de masse n'est pas valable que pour un certain type de réactions, y compris les réactions homogènes. On suppose au cours de ce travail que les réactions surfaciques entre aussi dans ce formalisme.

En supposant qu'aucun phénomène de précipitation ne se produise et en tenant compte de la distinction entre les espèces primaires et les espèces secondaires, il est possible de définir la loi d'action de masse sous la forme suivante :

$$\begin{cases} \{C_i\} = K_i^m \prod_{k=1}^{n_{pm}} \{X_k\}^{\mu_{1,i,k}} & i = 1, ..., n_{sm} \\ \{CS_i\} = K_i^s \prod_{k=1}^{n_{pm}} \{X_k\}^{\mu_{2,i,k}} \prod_{k=1}^{n_{pf}} \{S_k\}^{\mu_{3,i,k}} & i = 1, ..., n_{sf} \end{cases} . \tag{1.12}$$

où $\{C_i\}$ et $\{CS_i\}$ sont les activités de chaque espèce secondaire mobile C_i et fixée CS_i données par la loi d'action de masse par l'intermédiaire des activités $\{X_k\}$ et $\{S_k\}$ de chaque espèce primaire mobile X_k et fixée S_k. K^m est la constante d'équilibre pour les réactions entre les espèces mobiles et K^s est la constante d'équilibre pour les réactions de sorption.

Activité d'une espèce chimique Dans une solution, toute espèce chimique va interagir avec ses voisins ainsi qu'avec le solvant. La disponibilité de l'espèce chimique vis-à-vis d'une réaction peut alors apparaître assez différente de la concentration dans la solution. Cet écart est exprimé par la notion d'***activité***. Les lois d'action de masse (1.12) sont exprimées en fonction de l'activité des espèces primaires et secondaires. La relation (1.13) entre l'activité d'une espèce $\mathcal{E}_j$ et sa concentration est donnée par le cœfficient d'activité γ_j, calculé en utilisant des modèles spécifiques (Davies, debye-Huckel, etc.) :

$$\{\mathcal{E}_j\} = \gamma_j [\mathcal{E}_j]. \tag{1.13}$$

Dans cette thèse, nous ne considérerons que le cas d'une solution *idéale* - le cas où l'espèce ne subit aucune interaction et le coefficient d'activité γ est par suite égale à 1 et nous ne prendrons donc pas en compte les modèles de correction d'activité proposés dans la littérature [26]. Cela revient à confondre activité et concentration, et pour simplifier les notations, nous noterons donc par X_j, S_j, C_i ou CS_i

la concentration de l'espèce correspondante. La loi d'action de masse (1.12) se reécrit donc sous la forme :

$$\begin{cases} C_i = K_i^m \prod_{k=1}^{n_{pm}} X_k^{\mu_{1,i,k}} & i = 1, ..., n_{sm} \\ CS_i = K_i^s \prod_{k=1}^{n_{pm}} X_k^{\mu_{2,i,k}} \prod_{k=1}^{n_{pf}} S_k^{\mu_{3,i,k}} & i = 1, ..., n_{sf} \end{cases} . \tag{1.14}$$

1.2.4 Loi de conservation des espèces chimiques

Si on considère un système fermé, c'est à dire sans échange de matière avec l'extérieur, dont toutes les réactions chimiques sont des réactions à l'équilibre, alors on peut supposer que la quantité totale d'un espèce $\mathcal{X}_j$ dans le système est invariante. Ceci est exprimée en fonction d'une concentration totale T_j^m si cette quantité est présente sous forme d'espèce aqueuse considérée, et T_j^s si cette quantité est présente sous la forme d'espèce sorbée. Ainsi, la loi de conservation en concentration totale (connue encore sous le nom de Loi de Lavoisier) est donnée par les deux relations suivantes :

$$\begin{cases} T_j^m = X_j + \sum_{i=1}^{n_{sm}} \mu_{1,i,j} C_i + \sum_{i=1}^{n_{sf}} \mu_{2,i,j} CS_i & j = 1, ..., n_{pm} \\ T_j^s = S_j + \sum_{i=1}^{n_{sm}} \mu_{3,i,j} CS_i & j = 1, ..., n_{pf} \end{cases} \tag{1.15}$$

Les relations dans (1.15) introduisent une distinction entre la concentration en composant, mobile X_j et fixée S_j, et les autres concentrations des autres espèces secondaires C_i et CS_i. Or cette distinction n'est pas forcément nécessaire. On peut tout à fait considérer la réaction (1.16) pour chaque composant mobile ou fixée, dont le coefficient stoechiométrique est égale à 1 et la constante d'équilibre est égale à 1 :

$$\mathcal{X}_j \rightleftarrows \mathcal{C}_i. \tag{1.16}$$

La concentration totale en composant mobile X_j et la concentration totale en composant fixée S_j s'écrivent alors tout simplement :

$$\begin{cases} T_j^m = \sum_{i=1}^{n_{sm}} \mu_{1,i,j} C_i + \sum_{i=1}^{n_{sf}} \mu_{2,i,j} CS_i & j = 1, ..., n_{pm} \\ T_j^s = \sum_{i=1}^{n_{sm}} \mu_{3,i,j} CS_i & j = 1, ..., n_{pf} \end{cases} \tag{1.17}$$

ou sous la forme matricielle suivante :

$$\begin{cases} T^m = \mu_1^T . C + \mu_2^T . CS \\ T^s = \mu_3^T . CS \end{cases} \tag{1.18}$$

1.2.5 Tableau de Morel

Les différentes relations caractérisant un système chimique ; réactions chimiques (1.10)-(1.11), lois de conservation (1.17) et lois d'action de masse (1.14), peuvent être synthétiquement présentés sous forme de *Tableau des équilibres* ou *Tableau de Morel*, introduit par Morel [27]. Dans ce tableau, on écrit tout simplement la matrice μ, en indiquant au dessus de chaque colonne (resp. à gauche de chaque ligne) l'espèce primaire (resp. l'espèce secondaire) correspondante. En bas de chaque colonne se trouve la concentration totale de l'espèce considérée et à droite de chaque ligne se trouve la valeur de la constante d'équilibre de la réaction correspondante (cf. tableau 1.1).

TABLE 1.1 – Tableau de *Morel*

Espèces	X^T	S^T	**Constante d'équilibre** $\mathbf{K}^T$
C	μ_1	0	K^m
CS	μ_2	μ_3	K^s
Concentration totale $\mathbf{T}^T$	$(T^m)^T$	$(T^s)^T$	

La lecture du *Tableau de Morel* permet donc de connaître :

1. La réaction chimique, ou de manière équivalente, la loi d'action de masse pour chaque espèce secondaire peut être lue sur une ligne du tableau.
2. La loi de conservation d'une espèce primaire peut être lue sur une colonne.

Cette présentation permet, sur un exemple concret, de présenter en une seule fois l'ensemble des informations. En absence de phénomènes de précipitation, le *Tableau de Morel* pour un système chimique est donné, sous forme générale, par le tableau 1.1.

1.2.6 Système chimique

En substituant les lois d'action de masse données par (1.14) dans les équations de conservations données par (1.17), on peut présenter l'équilibre chimique par un système de $(n_{pm}+n_{pf})$ équations algébriques non linéaires à $(n_{pm}+n_{pf})$ inconnues, ne dépendant que des concentrations en composants :

$$\begin{cases} T_j^m = \sum_{i=1}^{n_{sm}} \mu_{1,i,j}\left(K_i^m \prod_{k=1}^{n_{pm}} X_k^{\mu_{1,i,k}}\right) + \sum_{i=1}^{n_{sf}} \mu_{2,i,j}\left(K_i^s \prod_{k=1}^{n_{pm}} X_k^{\mu_{2,i,k}} \prod_{k=1}^{n_{pf}} S_k^{\mu_{3,i,k}}\right) & j=1,...,n_{pm} \\ T_j^s = \sum_{i=1}^{n_{sm}} \mu_{3,i,j}\left(K_i^s \prod_{k=1}^{n_{pm}} X_k^{\mu_{2,i,k}} \prod_{k=1}^{n_{pf}} S_k^{\mu_{3,i,k}}\right) & j=1,...,n_{pf} \end{cases} \tag{1.19}$$

Le système (1.19) est un système d'équations algébrique non linéaire, donc devant l'impossibilité de mettre en œuvre une résolution directe (en général), il est nécessaire de faire appel à des méthodes itératives.
Avant de présenter la procédure de résolution de ce système et de discuter les méthodes numériques utilisées, il est important de donner les résultats théoriques existants dans la littérature, concernant l'existence et l'unicité de la solution, ainsi qu'il est nécessaire de modifier un peu le système afin de diminuer son niveau de non linéarité pour ne pas mettre en échec les algorithmes numériques qu'on va utiliser. On verra dans la suite que cette modification sera faite grâce à un changement de variable bien adapté.

1.2.6.1 Existence et unicité de la solution

La plupart des travaux discutant de cet objectif partent de la formulation sous forme de minimisation de l'enthalbie libre de la réaction ou l'Energie de Gibbs. Un des premiers articles dans cette veine est celui de Shapiro et Shapley [28]. Ces auteurs prouvent que l'existence d'au moins un minimum est établie sous des hypothèses assez générales (systèmes à plusieurs phases). Les lois de conservation des espèces primaires (1.17) constituent les contraintes du problème de minimisation, donc il est évidemment nécessaire qu'elles soient vérifiées, ce qui constitue une condition sur les concentrations totales T^m et T^s. La convexité de la fonctionnelle d'énergie implique l'existence d'une solution. D'autre part, l'unicité est bien vérifiée pour un système chimique à une seule phase, ce qui recouvre les cas traités dans cette thèse.

Une autre approche pour l'existence et l'unicité est donnée par G. Gnacadja [29] : la loi d'action de masse (1.14) constitue un système d'équations polynomiales, auxquel l'auteur applique un théorème de point fixe, pour une métrique adaptée au problème.

1.2.6.2 Changement de variables : Formulation en logarithme

Les inconnues du système (1.19) sont les concentrations des différentes espèces primaires. Ceci rend le système difficile à résoudre. En effet, ces concentrations sont susceptibles de varier sur plusieurs ordres de grandeur, et doivent rester positives pour garder leur signification physique. Ces deux contraintes rendent la résolution numérique délicate. Heureusement, par un simple changement de variables, on peut surmonter ces deux difficultés. Ce changement a été adopté par la majorité des codes de calcul : on prend comme inconnues les logarithmes des concentrations. Ainsi, les concentrations seront automatiquement positives et les inconnues du système non linéaire garderont un ordre de grandeur raisonnable.

On propose dans notre travail d'écrire le système chimique en logarithme de base 10 ($\log_{10}$) de la concentration des composants mobiles et fixes :

$$\xi_j = \log_{10}(X_j) \quad \text{et} \quad \eta_k = \log_{10}(S_k), \tag{1.20}$$

ou sous une formalisme matriciel :

$$\xi = \log_{10}(X) \quad \text{et} \quad \eta = \log_{10}(S). \tag{1.21}$$

Soient $\mathbf{K}^m = \log_{10}(K^m)$ et $\mathbf{K}^s = \log_{10}(K^s)$. Les conséquences de ce changement sur le système (1.19) sont limitées. On peut reformuler la loi d'action de masse comme suit :

$$\begin{cases} C_i = 10^{(\mathbf{K}_i^m + \sum_{k=1}^{npm} \mu_{1,i,k}\xi_k)} & i = 1, ..., n_{sm} \\ CS_i = 10^{(\mathbf{K}_i^s + \sum_{k=1}^{npm} \mu_{2,i,k}\xi_k + \sum_{k=1}^{npf} \mu_{3,i,k}\eta_k)} & i = 1, ..., n_{sf} \end{cases} \tag{1.22}$$

ou sous la forme matricielle :

$$\begin{cases} C = 10^{(\mathbf{K}^m + \mu_1 \times \xi)} \\ CS = 10^{(\mathbf{K}^s + \mu_2 \times \xi + \mu_3 \times \eta)} \end{cases} \tag{1.23}$$

Où "$\times$" désigne le symbole du produit matricielle.
Le système non linéaire (1.19) s'écrit ainsi en fonction des variables ξ et η :

$$\begin{cases} T_j^m = \sum_{i=1}^{n_{sm}} \mu_{1,i,j}.10^{(\mathbf{K}_i^m + \sum_{k=1}^{npm} \mu_{1,i,k}\xi_k)} + \sum_{i=1}^{n_{sf}} \mu_{2,i,j}.10^{(\mathbf{K}_i^s + \sum_{k=1}^{npm} \mu_{2,i,k}\xi_k + \sum_{k=1}^{npf} \mu_{3,i,k}\eta_k)} & j = 1, ..., n_{pm} \\ T_j^s = \sum_{i=1}^{n_{sm}} \mu_{3,i,j}.10^{(\mathbf{K}_i^s + \sum_{k=1}^{npm} \mu_{2,i,k}\xi_k + \sum_{k=1}^{npf} \mu_{3,i,k}\eta_k)} & j = 1, ..., n_{pf} \end{cases} \tag{1.24}$$

et sous la forme matricielle :

$$\begin{cases} T^m &= \mu_1^T \times 10^{(\mathbf{K}^m + \mu_1 \times \xi)} + \mu_2^T \times 10^{(\mathbf{K}^s + \mu_2 \times \xi + \mu_3 \times \eta)} \\ T^s &= \mu_3^T \times 10^{(\mathbf{K}^s + \mu_2 \times \xi + \mu_3 \times \eta)} \end{cases} \tag{1.25}$$

qui peut être réduite comme suit :

$$\mathbf{T} = \mu^T \times 10^{(\mathbf{K} + \mu \times \omega)}, \tag{1.26}$$

avec

$$\mathbf{T} = \begin{pmatrix} T^m \\ T^s \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} \mathbf{K}^m \\ \mathbf{K}^s \end{pmatrix}, \quad \omega = \begin{pmatrix} \xi \\ \eta \end{pmatrix} \quad \text{et} \quad \mu = \begin{pmatrix} \mu_1 & 0 \\ \mu_2 & \mu_3 \end{pmatrix}. \tag{1.27}$$

Le système non linéaire (1.25) constitue le *problème chimique* portant uniquement sur ξ et η, où Les vecteurs de concentrations totales des espèces composants, mobiles et fixés, T^m et T^s, sont données et bien connues. Outre la simplicité des notations, le changement de variables adapté fait apparaître clairement la nature linéaire des lois d'action de masse, et met en évidence la difficulté du problème à résoudre.

Une fois que la solution unique de ce problème, notée (ξ^*, η^*), est trouvée, on peut calculer les concentrations des espèces secondaires en utilisant la loi d'action de masse (1.22) (ou (1.23)).

1.2.6.3 Méthode des Fractions Continues Positives

Telle qu'elle est présentée dans les travaux de Wigley [31] et Parkhurst et Appelo [16], la méthode des fractions continues n'est pas adaptées au calcul d'équilibres thermodynamiques complexes. Sous sa forme générale, cette méthode s'est révélée trop rigide et ne permet pas la prise en compte de cœfficients stœchiométriques négatifs. La nouvelle méthode des Fractions continues positives (FCP), introduite par J. Carrayrou dans [22], permet de s'affranchir des difficultés qui limitent la méthodes des Fractions Continues. Il est alors possible de résoudre des problèmes de spéciation comprenant des cœfficients stœchiométriques négatifs ou des concentrations totales en composants non strictement positives.

Pour ce faire, nous définissons deux grandeurs analogues aux concentrations totales en espèces primaires, la *somme des réactifs* (1.28) et la *somme des produits* (1.29), construites de telle sorte qu'elles soient toujours positive.
Avec $\mathcal{C} = \begin{pmatrix} C \\ CS \end{pmatrix}$ et $n_p = n_{pm} + n_{pf}$, ces deux grandeurs sont définies comme suit :

$$\mathcal{S}_j^R = \begin{cases} \sum\limits_{\mu_{i,j}>0} \mu_{i,j}.\mathcal{C}_i & \text{si} \quad \mathbf{T}_j \geq 0 \\ |\mathbf{T}_j| + \sum\limits_{\mu_{i,j}>0} \mu_{i,j}.\mathcal{C}_i & \text{si} \quad \mathbf{T}_j < 0 \end{cases} \qquad j = 1, ..., n_p \tag{1.28}$$

$$\mathcal{S}_j^P = \begin{cases} \mathbf{T}_j + \sum\limits_{\mu_{i,j}<0} |\mu_{i,j}|.\mathcal{C}_i & \text{si} \quad \mathbf{T}_j \geq 0 \\ \sum\limits_{\mu_{i,j}<0} |\mu_{i,j}|.\mathcal{C}_i & \text{si} \quad \mathbf{T}_j < 0 \end{cases} \qquad j = 1, ..., n_p \tag{1.29}$$

En utilisant ces deux nouvelles valeurs, le bilan de masse en composant $\mathcal{X}_j$ (mobile ou fixé) s'écrit à l'équilibre comme égalité entre la somme des réactifs et la somme des produits :

$$\mathcal{S}_j^R = \mathcal{S}_j^P. \tag{1.30}$$

On choisit une espèce $\mathcal{E}_{i_0}$ pour laquelle le cœfficient stœchiométrique $\mu_{i_0,j}$ pour le composant $\mathcal{X}_j$ est non nul. Les lois d'action de masse sont écrites pour la somme des réactifs si $\mu_{i_0,j}$ est positif (respectivement, pour la somme des produits si $\mu_{i_0,j}$ est négatif) en utilisant les concentrations des composants aux itérations n et $n+1$. Donc en détaillant la relation (1.30), on obtient :

$$(\mathcal{X}_{j,n+1})^{\mu_{i_0,j}}.\Big[\sum_{\mu_{i,j}>0} \mu_{i,j}\mathbf{K}_i \prod_{k\neq j} (\mathcal{X}_{j,n})^{\mu_{i,k}}.(\mathcal{X}_{j,n})^{\mu_{i,j}-\mu_{i_0,j}}\Big] = \mathbf{T}_j + \sum_{\mu_{i,j}<0} |\mu_{i,j}|\mathcal{C}_{i,n} \tag{1.31}$$

avec $\mathcal{X} = \begin{pmatrix} X \\ S \end{pmatrix}$, $\mathcal{X}_{j,n}$ est la concentration du $j^{\text{ème}}$ composant $\mathcal{X}_j$ à l'itération n et $\mathcal{C}_{i,n}$ est celle de l'$i^{\text{ème}}$ espèce secondaire $\mathcal{C}_i$ à l'itération n.
En pratique, il est très souvent possible de choisir $\mathcal{E}_{i_0}$ identique à $\mathcal{X}_j$, ce qui implique que $\mu_{i_0,j} = 1$. Sinon, pour les autres cas, on propose de choisir $\mathcal{E}_{i_0}$ de sorte que $\mu_{i_0,j}$ soit le plus petit cœfficient stœchiométrique strictement positive dans la matrice μ. Après réarrangement, la relation (1.31) donne :

$$(\mathcal{X}_{j,n+1})^{\mu_{i_0,j}} = \frac{(\mathcal{X}_{j,n})^{\mu_{i_0,j}}}{(\mathcal{X}_{j,n})^{\mu_{i_0,j}}} \frac{\mathbf{T}_j + \sum_{\mu_{i,j}<0} |\mu_{i,j}| \mathcal{C}_{i,n}}{\sum_{\mu_{i,j}>0} \mu_{i,j} \mathbf{K}_i \prod_{k \neq j} (\mathcal{X}_{j,n})^{\mu_{i,k}} . (\mathcal{X}_{j,n})^{\mu_{i,j} - \mu_{i_0,j}}}. \tag{1.32}$$

Les sommes des réactifs et des produits apparaissent dans (1.32), qui s'écrit alors sous une forme donnant $\mathcal{X}_{j,n+1}$:

$$\mathcal{X}_{j,n+1} = \mathcal{X}_{j,n} \left(\frac{\mathcal{S}^P_{j,n}}{\mathcal{S}^R_{j,n}} \right)^{\frac{1}{\mu_{i_0,j}}}. \tag{1.33}$$

Rappelons maintenant qu'on a $\omega_{n+1} = \log_{10}(\mathcal{X}_{n+1})$ alors, écrite en fonction du logarithme des concentrations des espèces primaires, la relation (1.33) devient :

$$\omega_{j,n+1} = \omega_{j,n} + \frac{1}{\mu_{i_0,j}} \left[\log_{10}(\mathcal{S}^P_{j,n}) - \log_{10}(\mathcal{S}^R_{j,n}) \right]. \tag{1.34}$$

Cette nouvelle relation peut alors être considérée comme l'itération conventionnelle de point fixe donnée par :

$$\omega_{n+1} = \mathbf{G}(\omega_n) \qquad n = 0, 1, ... \tag{1.35}$$

où $\mathbf{G}$ est la fonction fabriquée et définie de la manière suivante :

$$\mathbf{G} : \begin{array}{ccc} \mathbb{R}^{n_p} & \longrightarrow & \mathbb{R}^{n_p} \\ \omega & \longmapsto & \mathbf{G}(\omega) \end{array} \quad \Big| \quad \mathbf{G}(\omega) = \omega + \frac{1}{\mu_0} \left[\log_{10}(\mathcal{S}^P) - \log_{10}(\mathcal{S}^R) \right]. \tag{1.36}$$

La résolution du problème d'équilibre thermodynamique (1.19) revient donc à résoudre le problème de point fixe donné par :

$$\omega = \mathbf{G}(\omega). \tag{1.37}$$

1.3 RÉSOLUTION DU PROBLÈME D'ÉQUILIBRE CHIMIQUE

L'objectif de cette partie est double :

1. Le premier est de présenter les méthodes itératives utilisées pour résoudre le problème d'équilibre thermodynamique (1.19) comme étant un problème de point fixe donné par (1.37) dont la solution est notée $\omega_* = \log_{10}(\mathcal{X}^*)$. Partant d'un vecteur $\omega_0 = \log_{10}(\mathcal{X}_0)$ approprié, où $\mathcal{X}_0$ est une approximation initiale de la solution $\mathcal{X}^*$, la séquence $\{\omega_n\}$ peut être générée par l'itération de point fixe (IPF) (1.35). Pour améliorer le taux de convergence de (IPF), on considère trois méthodes itératives : la méthode d'*Accélération d'Anderson* (AA) et les deux principales méthodes d'extrapolation vectorielle de type polynomial, *Minimal Polynomial Extrapolation* (MPE) et *Reduced Rank Extrapolation* (RRE). On donnera les définitions, les implémentations, les principaux algorithmes de ces méthodes ainsi que leur application au problème (1.37).
2. Le second objectif se base sur une étude expérimentale faite en mettant en œuvre les trois méthodes numériques citées. On présente les résultats numériques pour le problème d'équilibre thermodynamique sur le cas test d'Acide Gallique et sur le benchmark MoMas, cas test 1D "Easy". Tout d'abord, la description des cas tests est donnée. Dans un deuxième temps, pour le test d'acide gallique, les résultats sont comparés aux résultats de Jérôme Carrayrou issus des

méthodes de type Newton-Raphson, développés dans [29]. Pour le cas test 1D "Easy" de Benchmark MoMas, on compare nos résultats aux résultats obtenus par certains codes de transport réactifs qui ont participé à la réalisation du Benchmark. Parmi ces codes, on cite le code HYTEC [19], où toutes les réactions chimiques sont résolues par le code de spéciation CHESS [18] qui utilise un schéma amélioré de Newton-Raphson pour résoudre l'ensemble des équations algébriques non linéaires décrivant le système chimique. On donne également une comparaison de nos résultats avec ceux obtenus par d'autres codes [39] comme SPECY, MIN3P, GDAE et Hoffmann et al, qui sont basés sur une méthode de type Newton pour linéariser le système chimique et qui utilise chacun une méthode spécifique pour trouver la solution du système d'équations linéarisé. Ces comparaisons assurent la rapidité, la robustesse ainsi que la stabilité des méthodes (AA), (MPE) et (RRE) et montrent aussi et surtout qu'elles ont une meilleure performance en terme de temps CPU.

1.3.1 Méthode d'Accélération d'Anderson

Pour l'itération de point fixe (IPF) (1.35), la forme générale habituelle de l'Accélération d'Anderson [1, 32] est formulée par l'algorithme 1 où $\beta_k > 0$ est un paramètre de relaxation. Dans certains domaines d'application, tels que les calculs de structure électronique, β_k est appelé *Anderson mixing coefficient* et la méthode d'Accélération d'Anderson est appelée *Anderson mixing method*. Dans ce travail, comme dans [34], on ne considère que le cas où $\beta_k = 1$ dans l'algorithme 1.

D'après Toth et Kelley [34], la méthode AA (m), avec $\beta_k = 1$, converge lorsque l'application du point fixe $\mathbf{G}$ est une contraction. Ainsi, le taux de convergence n'est pas pire que celui dans l'itération de Picard (IPF) (1.35).
L'idée de l'algorithme 1 est de définir la $(k+1)^{\text{ième}}$ itération comme une combinaison des valeurs de $\mathbf{G}$, dans laquelle les cœfficients sont déterminés en minimisant la norme d'une combinaison linéaire de vecteurs résiduels. Dans cet algorithme, au plus $m+1$ vecteurs résiduels sont enregistrés à chaque itération. À la $k^{\text{ième}}$ itération, si $k < m$ alors le dernier vecteur résiduel f_k sera ajouté à la matrice F_k à droite ; sinon, pour $k \geq m$ alors le plus ancien vecteur résiduel f_{k-m_k} est également supprimé de F_k à gauche. À la $k^{\text{ième}}$ itération, $m_k + 1$ vecteurs sont enregistrés dans F_k ($m_k + 1 < m$). m_k est appelé

Algorithme 1 : ANDERSON ACCELERATION (AA-I)

Entrées : Le vecteur ω_0 et un entier $m \geq 1$.

Définir $\omega_1 = \mathbf{G}(\omega_0)$ et $f_0 = \mathbf{G}(\omega_0) - \omega_0$;

pour $k = 1, 2, \ldots$ **faire**

Définir $m_k = \min\{m, k\}$;
Calculer $\mathbf{G}(\omega_k)$ et laisser $f_k = f(\omega_k) = \mathbf{G}(\omega_k) - \omega_k$;
Définir $F_k = (f_{k-m_k}, \ldots, f_k)$;
Déterminer $\alpha^{(k)} = (\alpha_0^{(k)}, \ldots, \alpha_{m_k}^{(k)})^T$ qui résoud :

$$\begin{aligned} &\underset{\alpha=(\alpha_0,\ldots,\alpha_{m_k})^T}{\text{minimiser}} && ||F_k\alpha||_2 \\ &\text{sous la contrainte} && \sum_{i=0}^{m_k} \alpha_i = 1 \end{aligned} \tag{1.38}$$

Définir $\omega_{k+1} = (1-\beta_k)\sum_{i=0}^{m_k} \alpha_i^{(k)}(\omega_{k-m_k+i}) + \beta_k \sum_{i=0}^{m_k} \alpha_i^{(k)} \mathbf{G}(\omega_{k-m_k+i})$;

fin

la *profondeur d'Anderson*. La profondeur maximale est spécifiée par le paramètre m, et la méthode est

souvent désignée par AA (m). On remarque que si $m = 0$ alors AA(m) est elle même l'itération de point fixe (IPF) (1.35).

1.3.1.1 Forme du problème des moindres carrés

Dans la mise en œuvre pratique, le problème des moindres carrés avec contraintes (1.38) est souvent formulé comme un problème des moindres carrés sans contraintes qui lui est équivalent [34, 35] :

$$\min_{\gamma=(\gamma_0,...,\gamma_{m_k-1})^T} ||f_k - \mathcal{F}_k\gamma||_2 \tag{1.39}$$

où

$$\mathcal{F}_k = (\Delta f_{k-m_k}, ..., \Delta f_{k-1}) \tag{1.40}$$

avec $\Delta f_i = f_{i+1} - f_i$, $i = k - m_k, ..., k-1$. Les vecteurs de cœfficient des moindres carrés γ et α sont liés par :

$$\alpha_0 = \gamma_0, \quad \alpha_j = \gamma_j - \gamma_{j-1} \quad 1 \leq j \leq m_k - 1 \quad \text{et} \quad \alpha_{m_k} = 1 - \gamma_{m_k-1}$$

Si la solution de (1.39) est donnée par $\gamma^{(k)} = (\gamma_0^{(k)}, ..., \gamma_{m_k-1}^{(k)})^T$, alors l'itération suivante s'exprime sous la forme suivante :

$$\omega_{k+1} = \mathbf{G}(\omega_k) - \sum_{i=1}^{m_k-1} \gamma_i^{(k)}[\mathbf{G}(\omega_{k-m_k+i+1}) - \mathbf{G}(\omega_{k-m_k+i})] = \mathbf{G}(\omega_k) - \mathcal{G}_k\gamma^{(k)},$$

où

$$\mathcal{G}_k = (\Delta\mathbf{G}_{k-m_k}, ..., \Delta\mathbf{G}_{k-1}) \tag{1.41}$$

avec $\Delta\mathbf{G}_i = \mathbf{G}(\omega_{i+1}) - \mathbf{G}(\omega_i)$, $i = k - m_k, ..., k-1$. On donne alors dans l'algorithme 2 une version plus spécifique (AA-II) de l'algorithme d'Accélération d'Anderson.

Algorithme 2 : ANDERSON ACCELERATION (AA-II)

Entrées : Le vecteur ω_0 et un entier $m \geq 1$.

Définir $\omega_1 = \mathbf{G}(\omega_0)$ et $f_0 = \mathbf{G}(\omega_0) - \omega_0$;

pour $k = 1, 2, ...$ **faire**

- Définir $m_k = \min\{m, k\}$;
- Calculer $\mathbf{G}(\omega_k)$ et laisser $f(\omega_k) = \mathbf{G}(\omega_k) - \omega_k$;
- Mettre à jour $\mathcal{F}_k$ et $\mathcal{G}_k$ par (1.40) et (1.41) ;
- Déterminer $\gamma^{(k)} = (\gamma_0^{(k)}, ..., \gamma_{m_k-1}^{(k)})$ qui résoud (1.39) ;
- Définir $\omega_{k+1} = \mathbf{G}(\omega_k) - \mathcal{G}_k\gamma^{(k)}$;

fin

1.3.1.2 Résolution du problème des moindres carrés

L'implémentation KINSOL de la méthode d'Accélération d'Anderson suit l'approche décrite par Walker dans [33]. Le problème des moindres carrés (1.39) est résolu en effectuant la factorisation QR de la matrice $\mathcal{F}_k$ et en utilisant la substitution en arrière pour résoudre le système triangulaire supérieur $R_k\gamma = Q_k^T f_k$. Notons que lorsque $k < m$, la taille de $\mathcal{F}_k$ est $n \times k$, et un nouveau vecteur est ajouté à droite de cette matrice à chaque itération. Après la $m^{\text{ième}}$ itération, $\mathcal{F}_k$ reste de taille fixe, $n \times m$, mais à chaque itération, un vecteur colonne est supprimé à gauche tandis qu'un nouveau vecteur est ajouté à droite. Il est inefficace de factoriser $\mathcal{F}_k$ de nouveau à chaque étape, donc deux procédures d'assistance sont utilisées pour mettre à jour Q_k et R_k, basées sur la factorisation précédente.

Mise à jour des facteurs Q_k et R_k Deux procédures sont utilisées : QRAdd et QRDelete.
QRAdd prend en compte l'ajout d'un vecteur à droite de la matrice $\mathcal{F}_k$, alors que QRDelete prend en compte la suppression d'un vecteur à gauche.
QRAdd utilise le processus de Gram-Schmidt modifié pour orthonormaliser chaque nouveau Δf_{k-1} par rapport aux colonnes précédentes de Q_k. Le vecteur résultant devient la nouvelle colonne la plus à droite de Q_k. L'algorithme 3 décrit la procédure QRAdd.
QRDelete est la procédure de mise à jour pour supprimer la colonne à gauche de $\mathcal{F}_{k-1}$ lorsque $m_{k-1} =$

Algorithme 3 : PROCÉDURE QRADD

Entrées : $Q \in \mathbb{R}^{n\times m_k}$, $R \in \mathbb{R}^{m_k\times m_k}$ et Δf_{k-1}.

pour $j = 1, ..., m_k - 1$ **faire**
  Définir $R(j, m_k) = Q(:, j)^T * \Delta f_{k-1}$;
  Mettre à jour $\Delta f_{k-1} \leftarrow \Delta f_{k-1} - R(j, m_k) * Q(:, j)$;
fin
Définir $Q(:, m_k) \leftarrow \Delta f_{k-1}/||\Delta f_{k-1}||_2$ et $R(m_k, m_k) = ||\Delta f_{k-1}||_2$.

m. Son idée générale est la suivante : Si $\mathcal{F}_{k-1} = Q * R$, alors $\mathcal{F}_{k-1}(:, 2 : m) = Q * R(:, 2 : m)$, où $R(:, 2 : m) \in \mathbb{R}^{m\times(m-1)}$ est une matrice de Hessenberg supérieure. On determine $m \times m$ rotations de Givens $J_1, ..., J_{m-1}$ telles que $J_{m-1} * ... * J_1 * R(:, 2 : m) \in \mathbb{R}^{m\times(m-1)}$ est une matrice triangulaire supérieure avec une ligne inférieure entièrement nulle. Ensuite

$$\mathcal{F}_{k-1}(:, 2 : m) = Q * R(:, 2 : m) = Q * J_1^T * ... * J_{m-1}^T * J_{m-1} * ... * J_1 * R(:, 2 : m),$$

et donc on met à jour $Q \leftarrow Q * J_1^T * ... * J_{m-1}^T(:, 1 : m - 1)$ et $R \leftarrow J_{m-1} * ... * J_1 * R(1 : m - 1, 2 : m)$. Cette procédure de mise à jour est assez longue et apparaît finalement deux fois dans la version finale de l'algorithme d'AA [33]. Il convient donc de la décrire sous forme d'un sous-programme. En utilisant le logiciel MATLAB pour l'implémentation de la méthode d'Accélération d'Anderson, on peut suivre la forme de la fonction en MATLAB "qrdelete", qui accomplit la même chose que la procédure QRDelete. La séquence d'appel pour ce but est "[Q,R] = qrdelete(Q,R,1)" (le "1" indique que la première colonne doit être supprimée). Dans l'entrée de cette séquence, les matrices $Q \in \mathbb{R}^{n\times m}$ et $R \in \mathbb{R}^{m\times m}$ sont telles que $\mathcal{F}_{k-1} = Q * R$; dans la sortie, $Q \in \mathbb{R}^{n\times(m-1)}$ et $R \in \mathbb{R}^{(m-1)\times(m-1)}$ sont telles que $\mathcal{F}_{k-1}(:, 2 : m) = Q * R$. On remarque qu'on a besoin seulement des facteurs Q et R.

1.3.1.3 Contrôle du conditionnement de la matrice $\mathcal{F}_k$

Si la profondeur maximale d'Anderson m est très petite, alors l'historique d'itération retenu peut ne pas être suffisant pour accélérer suffisamment la convergence. Cependant, si m est très grande, alors la matrice $\mathcal{F}_k$ peut devenir si mal conditionnée que la solution du problème des moindres carrés est inexacte et la stabilité numérique de l'algorithme global est affectée négativement.

Dans ce travail, on utilise la stratégie donnée dans [32] pour surveiller le nombre de conditionnement de la matrice $\mathcal{F}_k$ noté $cond(\mathcal{F}_k)$ et, si nécessaire, pour modifier la matrice afin de réduire son conditionnement, comme suit : lorsque $cond(\mathcal{F}_k)$ est supérieur à une tolérance donnée, les colonnes les plus à gauche de $\mathcal{F}_k$ sont supprimées une par une jusqu'à ce que $cond(\mathcal{F}_k)$ soit inférieur à la tolérance donnée.

Notons que $cond(\mathcal{F}_k)$, pour la norme l_2, est juste égale à $cond(R_k)$ dans la factorisation QR de $\mathcal{F}_k$. Par conséquent, pour la stratégie de contrôle utilisée dans ce travail, il est seulement nécessaire de surveiller le nombre de conditionnement de R_k et de le maintenir inférieur à la tolérance donnée. Lorsque $cond(R_k)$ est supérieur à cette tolérance, la suppression de la colonne la plus à gauche de $\mathcal{F}_k$

implique la mise à jour des facteurs Q_k et R_k [33]. Le processus global de cette stratégie est décrit par l'algorithme 4.

Algorithme 4 : GESTION DU NOMBRE DE CONDITIONNEMENT

Entrées : La tolérence $tol > 0$ et la factorisation QR : $\mathcal{F}_k = Q_k R_k$.

tant que $cond(R_k) > tol$ **faire**
  Supprimer la dernière colonne à gauche de $\mathcal{F}_k$ et mettre à jour la factorisation QR : $\mathcal{F}_k = Q_k R_k$
fin

1.3.2 Méthodes d'extrapolation vectorielle de type polynomial MPE et RRE

Définition Soit k un entier positif tel que $k \leq N$ et soit $\{x_k\}_{k\in\mathbb{N}}$ une suite de vecteurs dans $\mathbb{R}^N$. On définit la première et la deuxième différence directe de x_k, Δ et Δ^2 respectivement, par :

$$\Delta x_k = x_{k+1} - x_k \quad \text{et} \quad \Delta^2 x_k = \Delta x_{k+1} - \Delta x_k, \quad k = 0, 1, ... \tag{1.42}$$

Quand elles sont appliquées à la séquence vectorielle $\{x_k\}$, les méthodes d'extrapolation vectorielle MPE et RRE produisent une approximation t_k de la limite ou l'anti-limite de $\{x_k\}_{k\in\mathbb{N}}$; voir [2]. Evidemment, t_k est différente pour chaque méthode.
Soit $\mathcal{T}_k$ cette transformation, définie comme suit

$$\mathcal{T}_k : \begin{array}{ccc} \mathbb{R}^N & \longrightarrow & \mathbb{R}^N \\ x_k & \longmapsto & \mathcal{T}_k(x_k) \end{array} \quad \Big| \quad \mathcal{T}_k(x_k) = t_{k,q} = x_q + \sum_{i=1}^{k} b_i^{(q)} g_i(q), \quad q \geq 0 \tag{1.43}$$

où les coefficients $b_i^{(q)}$ sont des scalaires et $(g_i(q))_q$ sont les séquences de vecteurs auxiliaires pour ces méthodes d'extrapolation données par

$$g_i(q) = \Delta x_{q+i-1}, \quad i = 1, ..., k, \quad q \geq 0$$

Notons par $\tilde{\mathcal{T}}_k$, la nouvelle transformation produite à partir de $\mathcal{T}_k$, par :

$$\tilde{t}_{k,q} = x_{q+1} + \sum_{i=1}^{k} b_i^{(q)} g_i(q+1), \quad q \geq 0 \tag{1.44}$$

Les coefficients $b_i^{(q)}$ sont les mêmes dans les deux expressions (1.43) et (1.44).
Deux autres expressions pour les deux approximations $t_{k,q}$ et $\tilde{t}_{k,q}$ peuvent être données par :

$$t_{k,q} = \sum_{j=0}^{k} \nu_j^{(k)} x_{q+j} \quad \text{et} \quad \tilde{t}_{k,q} = \sum_{j=0}^{k} \nu_j^{(k)} x_{q+j+1} \tag{1.45}$$

avec q et k définissent le premier terme et le nombre de termes de la suite respectivement. La suite $\nu_j^{(k)}$ vérifie

$$\sum_{j=0}^{k} \nu_j^{(k)} = 1 \quad \text{et} \quad \sum_{j=0}^{k} \tau_{i,j} \nu_j^{(k)} = 0, \quad i = 0, 1, ..., k-1, \tag{1.46}$$

les scalaires $\tau_{i,j}$ étant définis par :

$$\tau_{i,j} = (u_i, \Delta x_{q+j}), \tag{1.47}$$

où les $u_i \in \mathbb{R}^N$ définissent la méthode utilisée.
En utilisant (1.46), la transformation (1.45) peut être exprimée sous la forme d'un quotient de deux déterminants

$$t_{k,q} = \frac{\begin{vmatrix} x_q & x_{q+1} & \dots & x_{q+k} \\ \tau_{0,0} & \tau_{0,1} & \dots & \tau_{0,k} \\ \vdots & \vdots & & \vdots \\ \tau_{k-1,0} & \tau_{k-1,1} & \dots & \tau_{k-1,k} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \dots & 1 \\ \tau_{0,0} & \tau_{0,1} & \dots & \tau_{0,k} \\ \vdots & \vdots & & \vdots \\ \tau_{k-1,0} & \tau_{k-1,1} & \dots & \tau_{k-1,k} \end{vmatrix}} \tag{1.48}$$

Notons que le déterminant dans le numérateur de (1.48) est le vecteur obtenu en développant ce déterminant par rapport à sa première ligne par la règle classique. Alors que le déterminant dans le dénominateur est égal à $\det(U_{k,q}^T \Delta^2 S_{k,q}) \neq 0$ qui est supposé être non nul. Le calcul de l'approximation $t_{k,q}$ nécessite les valeurs des termes $x_q, x_{q+1}, ..., x_{q+k+1}$, et peut être réalisé en utilisant l'un des algorithmes proposés dans [10].
Soient $U_{k,q} = span\,\{u_1^{(q)}, ..., u_k^{(q)}\}$ et $\Delta^i S_{k,q} = [\Delta^i x_q, ..., \Delta^i x_{q+k-1}]$, $i = 1, 2$. En remplaçant, dans le numérateur et le dénominateur de (1.48), chaque colonne j, $j = k+1, k+2, ...$, par la différence avec la colonne $j-1$, on obtient l'expression suivante :

$$t_{k,q} = \frac{\begin{vmatrix} x_q & \Delta S_{k,q} \\ U_{k,q}^T \Delta x_q & U_{k,q}^T \Delta^2 S_{k,q} \end{vmatrix}}{\left| U_{k,q}^T \Delta^2 S_{k,q} \right|} \tag{1.49}$$

Avec ces notations et en utilisant la formule de Schur, $t_{k,q}$ peut s'exprimer sous la forme suivante :

$$t_{k,q} = x_q - \Delta S_{k,q} (U_{k,q}^T \Delta^2 S_{k,q})^{-1} U_{k,q}^T \Delta x_q. \tag{1.50}$$

D'après ce qui précède, nous avons le résultat suivant

Proposition 1 *Soit $t_{k,q}$ la transformation donnée par l'expression (1.50). Alors cette transformation existe si et seulement si $\det(U_{k,q}^T \Delta^2 S_{k,q}) \neq 0$.*

Pour des valeurs variables de k et q, le calcul de $t_{k,q}$ peut être effectué par certains des algorithmes proposés par Sidi et Ford dans [4].

Résidu généralisé K. Jbilou et H. Sadok ont réalisé de nombreux travaux sur les méthodes d'extrapolation MPE, RRE et MMPE [6, 7, 8]. Ils sont les premiers à avoir introduit des définitions des méthodes MPE, RRE et MMPE comme méthodes de projection en utilisant le résidu généralisé défini par :

$$\tilde{r}(t_{k,q}) = \tilde{t}_{k,q} - t_{k,q}. \tag{1.51}$$

D'après les relations (1.43) et (1.44), le résidu généralisé est donné par :

$$\tilde{r}(t_{k,q}) = \Delta x_q + \sum_{i=1}^{k} b_i^{(q)} g_i(q), \tag{1.52}$$

et en utilisant la relation (1.50), il peut être exprimé sous la forme suivante :

$$\tilde{r}(t_{k,q}) = \Delta x_q - \Delta^2 S_{k,q} (U_{k,q}^T \Delta^2 S_{k,q})^{-1} U_{k,q}^T \Delta x_q. \tag{1.53}$$

Notons que les coefficients $b_i^{(q)}$ impliqués dans (1.43) sont obtenus à partir de la relation d'orthogonalité :

$$\tilde{r}(t_{k,q}) \perp span\{u_1^{(q)}, u_2^{(q)}, ..., u_k^{(q)}\}. \tag{1.54}$$

Proposition 2 *Soient $\mathcal{W}_{k,q}$ et $\mathcal{U}_{k,q}$ les sous espaces définis par*

$$\mathcal{W}_{k,q} = span\{\Delta^2 x_q, ..., \Delta^2 x_{q+k-1}\} \quad et \quad \mathcal{U}_{k,q} = span\{u_1^{(q)}, ..., u_k^{(q)}\},$$

alors, d'après les relations (1.52) et (1.54), le résidu généralisé satisfait les deux conditions (i) et (ii)

(i) $\tilde{r}(t_{k,q}) - \Delta x_q \in \mathcal{W}_{k,q}$

(ii) $\tilde{r}(t_{k,q}) \perp \mathcal{U}_{k,q}$.

Ces deux dernières conditions montrent que le résidu généralisé $\tilde{r}(t_{k,q})$ est obtenu en projetant le vecteur Δx_q sur le sous-espace $\mathcal{W}_{k,q}$, orthogonalement à $\mathcal{U}_{k,q}$.

Implémentation Dans ce paragraphe et dans les sous-paragraphes "implémentation" suivants, pour une perspective d'implémentation, on s'intéressera seulement au cas où q est fixe. Sans restriction, on supposera même que $q = 0$ et on notera $t_{k,0} = t_k$ et $\Delta^i S_{k,0} = \Delta^i S_k$, pour $i = 1, 2$.
Le système linéaire (1.46) peut être exprimé sous la forme :

$$\begin{cases} \nu_0^{(k)} & + & \nu_1^{(k)} & + & \dots & + & \nu_k^{(k)} & = & 1 \\ \nu_0^{(k)}(u_0, \Delta x_0) & + & \nu_1^{(k)}(u_0, \Delta x_1) & + & \dots & + & \nu_k^{(k)}(u_0, \Delta x_k) & = & 0 \\ \nu_0^{(k)}(u_1, \Delta x_0) & + & \nu_1^{(k)}(u_1, \Delta x_1) & + & \dots & + & \nu_k^{(k)}(u_1, \Delta x_k) & = & 0 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ \nu_0^{(k)}(u_{k-1}, \Delta x_0) & + & \nu_1^{(k)}(u_{k-1}, \Delta x_1) & + & \dots & + & \nu_k^{(k)}(u_{k-1}, \Delta x_k) & = & 0 \end{cases} \tag{1.55}$$

Si on introduit les scalaires θ_i, pour $i = 0, ..., k$, definis par $\theta_i^{(k)} = \nu_i^{(k)} / \nu_k^{(k)}$, alors dans ce cas, on a

$$\nu_i^{(k)} = \frac{\theta_i^{(k)}}{\sum_{i=0}^{k} \theta_i^{(k)}} \quad i = 0, ..., k-1 \quad \text{et} \quad \theta_k^{(k)} = 1. \tag{1.56}$$

En utilisant ces nouvelles variables, le système (1.55) s'écrit sous la forme :

$$\begin{cases} \theta_0^{(k)}(u_0, \Delta x_0) & + & \theta_1^{(k)}(u_0, \Delta x_1) & + & \dots & + & \theta_{k-1}^{(k)}(u_0, \Delta x_{k-1}) & = & -(u_0, \Delta x_k) \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ \theta_0^{(k)}(u_{k-1}, \Delta x_0) & + & \theta_1^{(k)}(u_{k-1}, \Delta x_1) & + & \dots & + & \theta_{k-1}^{(k)}(u_{k-1}, \Delta x_{k-1}) & = & -(u_{k-1}, \Delta x_k) \end{cases}$$

Ce dernier système peut s'écrire encore sous la forme :

$$(U_k^T \Delta S_k)\theta^{(k)} = -\mathcal{U}_k^T \Delta x_k \tag{1.57}$$

où $\theta^{(k)} = (\theta_0^{(k)}, \dots, \theta_{k-1}^{(k)})^T$ et $\Delta S_k = (\Delta x_0, ..., \Delta x_{k-1})$.
On suppose maintenant que les coefficients $\nu_0^{(k)}, ..., \nu_k^{(k)}$ ont été calculés et on introduit les nouvelles variables suivantes :

$$\sigma_0^{(k)} = 1 - \nu_0^{(k)}, \quad \sigma_j^{(k)} = \sigma_{j-1}^{(k)} - \nu_j^{(k)} \quad j = 1, ..., k-1 \quad \text{et} \quad \sigma_{k-1}^{(k)} = \nu_k^{(k)}. \tag{1.58}$$

Alors, le vecteur t_k peut s'exprimer comme suit :

$$t_k = x_0 + \sum_{j=0}^{k-1} \sigma_j^{(k)} \Delta x_j = x_0 + \Delta S_k \sigma^{(k)} \tag{1.59}$$

où $\sigma = (\sigma_0, ..., \sigma_{k-1})^T$.
On remarque que pour déterminer les inconnus $\nu_i^{(k)}$, il faut tout d'abord calculer les $\theta_i^{(k)}$ en résolvant le système d'équations linéaires (1.57).
Notons qu'en utilisant les relations (1.51) et (1.59), le résidu général $\tilde{r}(t_k)$ peut être exprimé sous la forme suivante :

$$\tilde{r}(t_k) = \sum_{i=0}^{k} \nu_i^{(k)} \Delta x_i = \Delta S_{k+1} \nu^{(k)}$$

1.3.2.1 La méthode RRE

Définition La méthode RRE a été proposé, dans les années 1970, par Eddy [13] et Mesina [14]. Ils définient les scalaires $\tau_{i,j}$ de la manière suivante :

$$\tau_{i,j} = (\Delta^2 x_{q+i}, \Delta x_{q+j}). \tag{1.60}$$

$\mathcal{U}_{k,q}$ est donc définie par :

$$\mathcal{U}_{k,q} = \Delta^2 S_{k,q}, \tag{1.61}$$

et la transformation de la méthode RRE peut être exprimée, d'après la relation (1.50), sous la forme :

$$t_{k,q}^{\text{RRE}} = x_q - \Delta S_{k,q} (\Delta^2 S_{k,q}^T \Delta^2 S_{k,q})^{-1} \Delta^2 S_{k,q}^T \Delta x_q. \tag{1.62}$$

Implémentation On résume l'implémentation donnée par Sidi dans [4]. La méthode RRE est caractérisée par la résolution des problèmes des moindres carrés par la factorisation QR.
On introduit les notations suivantes :

$$\Delta S_{k+1} = (\Delta x_0, ..., \Delta x_k) \quad \text{et} \quad \nu^{(k)} = (\nu_0^{(k)}, ..., \nu_k^{(k)})^T. \tag{1.63}$$

En utilisant la formulation de l'équation (1.45) ainsi que le système linéaire (1.46), les $\nu_j^{(k)}$ peuvent être obtenues en résolvant le système de moindres carrées suivant :

$$\Delta S_{k+1} \nu^{(k)} = 0, \tag{1.64}$$

sous la contrainte

$$\sum_{j=0}^{k} \nu_j^{(k)} = 1. \tag{1.65}$$

Cela mène à minimiser la forme quadratique définie positive donnée par

$$(\nu^{(k)})^T \Delta S_{k+1}^T \Delta S_{k+1} \nu^{(k)},$$

sous la contrainte (1.65). D'après [5], les $\nu_i^{(k)}$ $(i = 0, 1, ..., k)$ peuvent être obtenues en résolvant le système de $(k+2)$ équations linéaires :

$$\Delta S_{k+1}^T \Delta S_{k+1} \nu^{(k)} = \lambda e, \sum_{j=0}^{k} \nu_j^{(k)} = 1 \tag{1.66}$$

où $e = (1, ..., 1)^T \in \mathbb{R}^{q+1}$ et λ est un scalaire strictement positive qui vérifie

$$\lambda = (\nu^{(k)})^T \Delta S_{k+1}^T \Delta S_{k+1} \nu^{(k)}. \tag{1.67}$$

Posons maintenant $d^{(k)} = (d_0^{(k)}, ..., d_k^{(k)})^T$ et $\lambda = (\sum_{i=0}^{k} d_i^{(k)})^{-1}$, alors $\nu^{(k)}$ peut être calculé en résolvant le système d'équations linéaires :

$$\Delta S_{k+1}^T \Delta S_{k+1} d^{(k)} = e. \tag{1.68}$$

On obtient

$$\nu^{(k)} = \lambda d^{(k)}. \tag{1.69}$$

Supposons que la matrice ΔS_{k+1} est de rang maximal, c'est à dire que $rang\ (\Delta S_{k+1}) = k+1$. Alors on peut définir la factorization QR de la matrice ΔS_{k+1} : $\Delta S_{k+1} = Q_k R_k$. Cela mène à écrire le système linéaire (1.68) sous la forme suivante :

$$R_k^T R_k d^{(k)} = e.$$

Finallement, on peut exprimer l'approximation t_k comme suit :

$$t_k^{\text{RRE}} = x_0 + Q_{k-1}(R_{k-1}\sigma^{(k)})$$

où $\sigma^{(k)} = (\sigma_0^{(k)}, \sigma_1^{(k)}, ..., \sigma_{k-1}^{(k)})^T$.
Une autre expression de t_k est donnée par :

$$t_k^{\text{RRE}} = x_0 + \sum_{j=0}^{k-1} \sigma_j^{(k)} \Delta x_j = x_0 + \Delta S_k \sigma^{(k)}. \tag{1.70}$$

Alors, en utilisant les relations (1.51) et (1.70), le résidu généralisé $\tilde{r}(t_k)$ peut s'écrire sous la forme :

$$\tilde{r}(t_k^{\text{RRE}}) = \sum_{i=0}^{k} \nu_i^{(k)} \Delta x_i = \Delta S_{k+1} \nu^{(k)}.$$

Notons que la factorisation QR de ΔS_{k+1} est formée en ajoutant une colonne supplémentaire à Q_{k-1} pour obtenir Q_k, et une colonne correspondante à R_{k-1} pour obtenir R_k. Cette factorisation peut être calculée avec peu de coût en appliquant le processus de Gram-Schimdt modifié (MGS) aux vecteurs $x_0, x_1, ..., x_{k+1}$; see [5].
Soient les matrices Q_k et R_k données par :

$$Q_k = (q_0|q_1|...|q_k) \quad \text{et} \quad R_k = \begin{pmatrix} r_{00} & r_{01} & r_{02} & \dots & r_{0k} \\ & r_{11} & r_{12} & \dots & r_{1k} \\ & & r_{22} & \dots & r_{2k} \\ & & & \ddots & \vdots \\ & & & & r_{kk} \end{pmatrix},$$

et soit $v_i = \Delta x_i$. Alors le processus (MGS) est résumé par l'algorithme 5.

D'après l'implémentation présentée ci-dessus, la méthode RRE est donnée par l'algorithme 6.

1.3.2.2 La méthode MPE

Définition La méthode MPE a été proposé par Cabay et Jackson dans [12]. Les scalaires $\tau_{i,j}$ sont définis par :

$$\tau_{i,j} = (\Delta x_{q+i}, \Delta x_{q+j}). \tag{1.71}$$

$\mathcal{U}_{k,q}$ est donc donné par :

$$\mathcal{U}_{k,q} = \Delta S_{k,q}, \tag{1.72}$$

et la tranformation de la méthode MPE est déduite de la relation (1.50) sous la forme suivante :

$$t_{k,q}^{\text{MPE}} = x_q - \Delta S_{k,q} (\Delta S_{k,q}^T \Delta^2 S_{k,q})^{-1} \Delta S_{k,q}^T \Delta x_q. \tag{1.73}$$

Algorithme 5 : MÉTHODE MGS

$r_{00} = ||v_0||_2$;
$q_0 = v_0/r_{00}$;
pour $k = 1, 2, ...$ **faire**
  $v_k^{(0)} = v_k$;
  pour $j = 0, ..., k-1$ **faire**
    $r_{jk} = (q_j, v_k^{(j)})$;
    $v_k^{(j+1)} = v_k^{(j)} - r_{jk} q_j$;
  fin
  $r_{kk} = ||v_k^{(k)}||_2$;
  $q_k = v_k^{(k)}/r_{kk}$;
fin

Algorithme 6 : MÉTHODE RRE

Entrées : Les vecteurs $x_0, x_1, ..., x_{k+1}$.

1. Calculer $v_i = \Delta x_i = x_{i+1} - x_i, \ \ i = 0, 1, ..., k$;
Poser $V_j = [v_0|v_1|...|v_{j-1}], \ \ j = 0, 1, ...$;
Calculer la factorisation QR de V_{k+1}, notée $V_{k+1} = Q_k R_k$;
($V_k = Q_{k-1} R_{k-1}$ est contenu $V_{k+1} = Q_k R_k$) ;

2. Résoudre le système linéaire

$$R_k^T R_k d^{(k)} = e \qquad d^{(k)} = [d_0^{(k)}, d_1^{(k)}, ..., d_k^{(k)}]^T \qquad e = [1, 1..., 1]^T$$

(Cela revient à résoudre deux systèmes triangulaires supérieur et inférieur) ;
Poser $\lambda = (\sum_{i=0}^{k} d_i^{(k)})^{-1}, \quad \lambda \in \mathbb{R}^+$;
Poser $\nu_i^{(k)} = \lambda d_i^{(k)}, \quad i = 0, 1, ..., k$;

3. Calculer $\sigma^{(k)} = [\sigma_0^{(k)}, \sigma_1^{(k)}, ..., \sigma_{k-1}^{(k)}]^T$ via :

$$\sigma_0^{(k)} = 1 - \nu_0^{(k)} \quad \text{et} \quad \sigma_j^{(k)} = \sigma_{j-1}^{(k)} - \nu_j^{(k)}, \quad j = 1, ..., k-1$$

.
Calculer t_k^{RRE} via :

$$t_k^{\text{RRE}} = x_0 + Q_{k-1}(R_{k-1}\sigma^{(k)})$$

.

Implémentation On suppose que la matrice ΔS_{k+1} est de rang maximal ($rang\ (\Delta S_{k+1}) = k+1$) alors il existe une factorisation QR de cette matrice : $\Delta S_{k+1} = Q_{k+1} R_{k+1}$, où $Q_{k+1} = (q_0|q_1|...|q_k) \in \mathbb{R}^{N \times k}$ est une matrice orthogonale et $R_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ est une matrice triangulaire supérieure dont les éléments diagonaux sont positifs. La matrice Q_{k+1} est obtenue à partir de la matrice $Q_k \in \mathbb{R}^{N \times k}$ en ajoutant le vecteur colonne q_k. De même, R_{k+1} est obtenue de $R_k \in \mathbb{R}^{k \times k}$ en ajoutant une ligne et une colonne à R_k. Cette factorisation QR peut être calculée avec peu de coût en appliquant le processus de Gram-Schmidt modifiée (MGS) aux vecteurs $x_0, x_1, ..., x_{k+1}$; voir l'algorithme 5.
On note par r_k le vecteur de $\mathbb{R}^k$ formé par la dernière colonne de la matrice R_{k+1} en enlevant le dernier élément, et par ρ_k un scalaire correspondant à cet élément. Alors cette factorisation QR peut s'écrire

sous la forme :

$$(\Delta S_k, \Delta x_k) = (Q_k, q_k) \begin{pmatrix} R_k & r_k \\ 0 & \rho_k \end{pmatrix} \tag{1.74}$$

En développant le membre à droite de l'équation (1.74), on obtient :

$$(\Delta S_k, \Delta x_k) = (Q_k R_k, Q_k r_k + \rho_k q_k). \tag{1.75}$$

En considérant la dernière colonne dans chacun des membres de (1.75), alors

$$\Delta S_k = Q_k r_k + \rho_k q_k. \tag{1.76}$$

Puisque $\Delta S_k = Q_k R_k$, alors en multipliant chaque membre de l'équation (1.76) par ΔS_k, on obtient :

$$\Delta S_k^T \Delta x_k = R_k^T Q_k^T (Q_k r_k + \rho_k q_k). \tag{1.77}$$

Puisque la matrice Q_{k+1} est orthogonale, alors en développant l'expression (1.77) à droite, on obtient la relation suivante :

$$\Delta S_k^T \Delta x_k = R_k^T r_k. \tag{1.78}$$

Par conséquent, le système linéaire (1.57) peut se simplifier de la manière suivante :

$$\begin{aligned} & (\Delta S_k^T \Delta x_k)\theta^{(k)} = -\Delta S_k^T \Delta x_k \\ \Leftrightarrow\ & R_k^T Q_k^T Q_k R_k \theta^{(k)} = -R_k^T r_k \\ \Leftrightarrow\ & R_k^T R_k \theta^{(k)} = -R_k^T r_k \\ \Leftrightarrow\ & R_k \theta^{(k)} = -r_k \quad (\text{Comme } R_k \text{ est non singulière}). \end{aligned}$$

Comme la matrice du système linéaire est triangulaire supérieure, la solution peut être facilement calculée par une méthode de remontée. Après le calcul de $\theta^{(k)}$, on calcule $\nu_i^{(k)}$, $i = 0, \ldots, k$ d'après (1.56) et $\sigma_i^{(k)}$, $i = 0, \ldots, k-1$ d'après (1.58). Finallement, on calcule l'approximation t_k^{MPE} sous la forme :

$$t_k^{\text{MPE}} = x_0 + Q_k R_k \sigma^{(k)}. \tag{1.79}$$

En se basant sur cette implémentation, la méthode MPE est donnée par l'algorithme 7.

1.3.2.3 Méthodes redémarrées (ou cycliques)

Quand on applique les algorithmes des méthodes RRE et MPE (algorithmes 7 et 6) dans leurs formes complètes, ils deviennent très coûteux lorsque k augmente, car le nombre de calculs requis augmente de façon quadratique avec le nombre d'itérations k. De plus, le coût de stockage augmente linéairement. Pour éviter cela et garder le coût de stockage et le coût des calculs les plus moins possible, ces algorithmes doivent être redémarrés périodiquement toutes les c étapes, pour certains entiers $c > 1$. L'algorithme 8 décrit la stratégie pratique de la méthode redémarrée.

1.3.2.4 Application des méthodes MPE et RRE à la résolution de système d'équilibre thermodynamique

Pour bien comprendre en détails comment se comportent les méthodes d'extrapolation MPE et RRE lorsqu'elles sont appliquées à la résolution de systèmes non linéaires et linéaires, on peut recourir aux travaux [5] et [6] de A. Sidi et K. Jbilou. Nous allons appliquer ces deux méthodes, en mode cyclique, au problème d'equilibre thermodynamique, on considèrera alors le système d'equations algébriques non linéaires (1.37).

Algorithme 7 : MÉTHODE MPE

Entrées : Les vecteurs $x_0, x_1, ..., x_{k+1}$.

1. Calculer $v_i = \Delta x_i = x_{i+1} - x_i,\ \ i = 0, 1, ..., k$;
Poser $V_j = [v_0|v_1|...|v_{j-1}],\ \ j = 0, 1, ...$;
Calculer la factorisation QR de V_{k+1}, notée $V_{k+1} = Q_{k+1}R_{k+1}$;
($V_k = Q_k R_k$ est contenu $V_{k+1} = Q_{k+1}R_{k+1}$) ;

2. Résoudre le système linéaire triangulaire supérieur

$$R_k d^{(k)} = -r_k \qquad d^{(k)} = [d_0^{(k)}, d_1^{(k)}, ..., d_{k-1}^{(k)}]^T \qquad r_k = [r_{0k}, r_{1k}, ...r_{(k-1)k}]^T$$

Poser $d_k^{(k)} = 1$ et calculer $\lambda = (\sum_{i=0}^{k} d_i^{(k)})^{-1}, \quad \lambda \in \mathbb{R}^+$;
Poser $\nu_i^{(k)} = \lambda d_i^{(k)}, \quad i = 0, 1, ..., k$;

3. Calculer $\sigma^{(k)} = [\sigma_0^{(k)}, \sigma_1^{(k)}, ..., \sigma_{k-1}^{(k)}]^T$ via :

$$\sigma_0^{(k)} = 1 - \nu_0^{(k)} \quad \text{et} \quad \sigma_j^{(k)} = \sigma_{j-1}^{(k)} - \nu_j^{(k)}, \quad j = 1, ..., k-1$$

.
Calculer t_k^{MPE} via :

$$t_k^{\text{MPE}} = x_0 + Q_k(R_k \sigma^{(k)})$$

.

Algorithme 8 : MÉTHODE REDÉMARRÉE TOUTES LES c ITÉRATIONS

Entrées : Pour $k = 0$, choisir un entier c et un vecteur x_0.

pour $k = 1, 2, ...$ **faire**
  Calculer les vecteurs $x_1, ..., x_c$;
  Calculer l'approximation t_{c-1} en utilisant l'algorithme correspondant à la méthode choisie ;
  si t_{c-1} *est satisfaisante*
    stop;
  sinon
    Poser $x_0 = t_{c-1}$;
  finsi
fin

Pour un vecteur arbitraire de logarithme de concentrations ω, le résidu est défini par

$$r(\omega) = \mathbf{G}(\omega) - \omega.$$

Partant d'un vecteur initial ω_0, une séquence $\{\omega_n\}$ est construite par l'itération de point fixe (1.35). On peut remarquer que

$$r(\omega_n) = \tilde{r}(\omega_n) = \Delta\omega_n \quad n = 0, 1, ...$$

En général, les méthodes d'extrapolation MPE et RRE sont plus efficaces si elles sont appliquées à des systèmes non linéaires préconditionnés où la nouvelle application de point fixe est obtenue à partir de l'ancienne par une technique de préconditionnement non linéaire. Afin de rendre le processus d'extrapolation plus efficace avec une stabilité numérique élevée, on propose de remplacer l'itération

de point fixe (1.35) par :

$$\omega_{n+1} = \tilde{\mathbf{G}}(\omega_n) \quad n = 0, 1, ..., \tag{1.80}$$

où

$$\tilde{\mathbf{G}}(\omega) = \omega + \kappa(\mathbf{G}(\omega) - \omega). \tag{1.81}$$

κ est un scalaire différent de 1 (pour $\kappa = 1$, la séquence générée est celle donnée par (1.35)). Ainsi ω_{n+1} est maintenant le pondéré "en moyenne" de ω_n et $\mathbf{G}(\omega_n)$, dans lequel les poids $1-\kappa$ et κ n'ont pas besoin d'être tous les deux positifs. En choisissant κ de manière appropriée, on peut faire de sorte que le spectre de la matrice jacobienne de $(1-\kappa)\omega + \kappa\mathbf{G}(\omega)$ en $\omega = \omega_*$ soit de plus en plus favorable à $t_{k,q}$ pour les grandes valeurs de q. Pour plus de détails, voir [5].

La séquence $\omega_1, \omega_2, \omega_3, ...$ est donc générée par l'itération de point fixe (1.80) et pour appliquer les méthodes d'extrapolation MPE et RRE au problème d'équilibre thermodynamique, on peut considérer l'algorithme 9.

Algorithme 9 : EXTRAPOLATION POUR LE PROBLÈME NON LINÉAIRE D'ÉQUILIBRE CHIMIQUE

Données : Le vecteur des concentrations totales des composants $\mathbf{T}$ et le vecteur des constantes d'équilibre $\mathbf{K}$.
Entrées : Pour $k = 0$, choisir ω_0 et les entiers p et l.

1. *Itération de base* :
$t_0 = \omega_0$;
$h_0 = t_0$;
$h_{j+1} = \tilde{\mathbf{G}}(h_j), \quad j = 0, ..., p-1$;

2. *Phase d'extrapolation* :
$s_0 = h_p$;
si $||s_1 - s_0|| < \epsilon$
  stop;
sinon
  $s_{j+1} = \tilde{\mathbf{G}}(s_j), \quad j = 0, ..., l$;
finsi
Calculer l'approximation t_l par RRE ou MPE ;

3. Définir $\omega_0 = t_l$, $k = k+1$ et retourner à **1.**.

Remarque 1 *De manière similaire aux problèmes linéaires [5], il est plus utile d'exécuter quelques itérations de base avant d'appliquer l'une des méthodes d'extrapolation pour résoudre le problème non linéaire (1.37). Deux manières peuvent être suivies pour ce but :*

1. *On peut exécuter N_0 itérations de base avant de démarrer le cycle, c'est-à-dire avant que MPE ou RRE soit appliqué pour la première fois ($N_0 \in \mathbb{N}$ se réfère à la taille de l'extrapolation) ;*
2. *On peut exécuter N itérations de base avant l'application de MPE ou RRE dans chaque cycle après le premier cycle.*

1.3.3 Résultats numériques

Les programmes spécifiques aux méthodes itératives (AA), (MPE) et (RRE) appliquées au problème d'équilibre thermodynamique (1.37) (ou (1.81)) ont été réalisés en utilisant le logiciel Matlab R2018a (voir Annexe A). Pour la méthode AA (algorithmes 1 et 2), on utilise le code MATLAB donné dans [33].

Pour la mise en œuvre des MPE et RRE (algorithmes 7 et 6), on convertit le programme informatique fourni dans [5] du langage Fortran en langage Matlab.
Quelques paramètres numériques essentiels pour les cas test sont :

- *En utilisant la méthode d'Accélération d'Anderson* :
  - on termine l'itération lorsque la norme résiduelle devient inférieure à 10^{-10} ;
  - la stratégie de surveillance du nombre de conditionnement est utilisée avec un seuil de suppression des colonnes $droptol = 10^{10}$;
  - le nombre maximal d'itérations non linéaires autorisé est $Kmax = 200$ itérations ;
- *En utilisant les deux méthodes MPE et RRE* :
  - le nombre maximum de cycles autorisé est $Ncycle = 30$;
  - la limite supérieure de $resc/resp$ utilisée comme critère d'arrêt est $epsc = 10^{-10}$, avec $resp$ est la norme l_2 du résidu pour $t(Kmax, N0)$ à la fin du premier cycle et $resc$ est la norme l_2 du résidu pour t à la fin de chaque cycle, récupéré à la fin du cycle suivant. Si $resc \leq epsc \times resp$ à la fin de certains cycles, alors un cycle supplémentaire est effectué et le $t(N, Kmax)$ correspondant est accepté comme approximation finale.
  - la limite supérieure de $res/R(0,0)$, le résidu relatif pour t, utilisé comme critère d'arrêt est $eps = 0$. Notons que $R(0,0)$ est la norme l_2 du résidu du vecteur initial ω_0. Si pour certains k, $res \leq eps \times R(0,0)$, alors le $t(0,k)$ correspondant est accepté comme approximation finale.

1.3.3.1 Cas test d' Acide Gallique

Présentation du test Il s'agit du test le plus simple, un système proposé par Brassard et Bodurtha (2000) [36] pour mettre en evidence des problèmes de convergence pour certaines méthodes numériques. Ce système est caractérisé par la présence de 17 espèces chimiques qui peuvent être décrites par la combinaison de 3 espèces primaires mobiles ($n_e = 17,\ n_{sm} = 14,\ n_{pm} = 3$), modélisant la spéciation de l'acide gallique en présence d'aluminium. Toutes les réactions décrivant ce système chimique sont homogènes entre les espèces mobiles ($n_{pf} = n_{sf} = 0$). Le pH est imposé à 5,8, ce qui donne un problème à deux inconnues : les concentrations des deux composants libres Al^{3+} et H_3L.

Le système chimique étudié est présenté dans le tableau 1.2 où les concentrations initiales de Al^{3+} et H_3L sont variables et les valeurs thermodynamiques sont proposées par Brassard et Bodurtha (2000) [36]. En fixant le pH du système, on remarque que la matrice des coefficients stœchiométriques se réduit à une matrice dont tous les coefficients sont positifs, de plus les concentrations totales en Al^{3+} et H_3L sont positives, alors la relation (1.81) s'écrit dans ce cas sous la forme :

$$\tilde{\mathbf{G}}(\xi) = \xi + \kappa \frac{I_{n_{pm}}}{\mu_{10}} . \left[\log_{10}(T^m) - \log_{10}(\mu_1^T . 10^{\mathbf{K}^m + \mu_1 \xi}) \right]. \tag{1.82}$$

Pour ce système chimique, on teste les méthodes itératives citées ci-dessus dans deux cas :

1 Concentrations initiales : $[Al^{3+}]_0 = 10^{-11}$ M ; $[H_3L]_0 = 5 \times 10^{-4}$M ;
2 Concentrations initiales : $[Al^{3+}]_0 = 5.012 \times 10^{-10}$ M ; $[H_3L]_0 = 10^{-9}$ M.

Résultats de la méthode AA Avec la méthode AA, le système d'équilibre chimique associé au cas test d'acide gallique (1.82) est étudié dans un premier temps sans relaxation ($\kappa = 1$).

1. **Convergence vers la solution**
   Pour les deux cas de concentrations initiales définis ci-dessus, on remarque la convergence de

TABLE 1.2 – Tableau des équilibres pour le test de l'acide gallique dont le pH est fixé à 5.8 ([22, 36]).

Espèces	H^+	Al^{3+}	H_3L	$\mathbf{K}^m$	$\mathcal{C}$ (**À l'équilibre**)
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_3^{6-}	-9	1	3	-37.69	1.14×10^{-10}
$Al_2(OH)_2(HL)_3^{2-}$	-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}
$Al_3(OH)_4(H_2L)^{4+}$	-5	3	1	-12.52	6.51×10^{-5}
Totale T (M)	pH=5.8	10^{-3}	10^{-3}		
Initiale $\mathcal{X}_0$ (M)	1.58×10^{-6}	variable	variable		
À l'équilibre $\mathcal{X}^*$ (M)	1.58×10^{-6}	2.03×10^{-5}	2.59×10^{-7}		

la méthode AA(m), pour toute profondeur maximale $m \geq 1$ (cf. Figure 1.1). AA($m = 1$) nécessite 26 itérations dans le cas 1 et 109 itérations dans le cas 2, tandis que AA($m \geq 2$) nécessite 16 itérations dans le cas 1 et 15 itérations dans le cas 2. Les premières itérations effectuées présentent des perturbations en termes de variation de concentrations $[Al^{3+}]$ et $[H_3L]$, mais elles ne sont plus compliquées et la convergence est obtenue sans difficulté. Ces perturbations résultent principalement du choix de la concentration initiale de chaque composant. Notons que la solution du système obtenue est égale à la solution de référence donnée par J. Carrayrou [22] et citée dans le tableau 1.2 :

$$\omega^* = \log_{10}(\mathcal{X}^*) \quad \text{avec} \quad \mathcal{X}^* = ([Al^{3+}]^*, [H_3L]^*)^T = (2.028 \times 10^{-5}, 2.6 \times 10^{-7})^T.$$

2. **Conditionnement de $\mathcal{F}_k$**
Grâce à la stratégie de surveillance du nombre de conditionnement utilisée, la méthode d'Accélération d'Anderson présente une grande stabilité et robustesse en résolvant le système d'équilibre chimique du cas test d'acide gallique. Le comportement du conditionnement de la matrice $\mathcal{F}_k$ est décrit dans la figure 1.2. Pour $m \in \{1, 2\}$, $cond\ (\mathcal{F}_k)$ reste inférieur à 10^{10} pour tout $k > 0$. Pour $m \geq 3$, cette matrice devient mal conditionnée à partir de la troisième itération ($cond\ (\mathcal{F}_{k\geq 3}) \geq 10^{15}$). Cependant, la stratégie de surveillance du conditionnement permet de faire baisser $cond(\mathcal{F}_k)$ en dessous de 10^{10} dès la troisième étape d'itération.

3. **Taux de convergence**
La Figure 1.3 montre que l'accélération de la convergence avec la méthode AA($m, m \geq 2$) com-

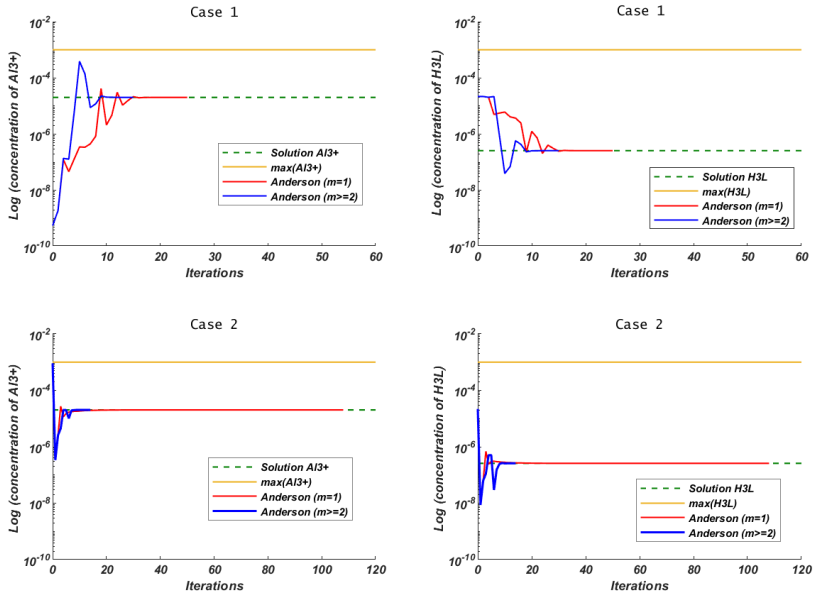

FIGURE 1.1 – Cas test d'acide gallique : Équilibre thermodynamique pour les composants H_3L et Al^{3+} par la méthode AA

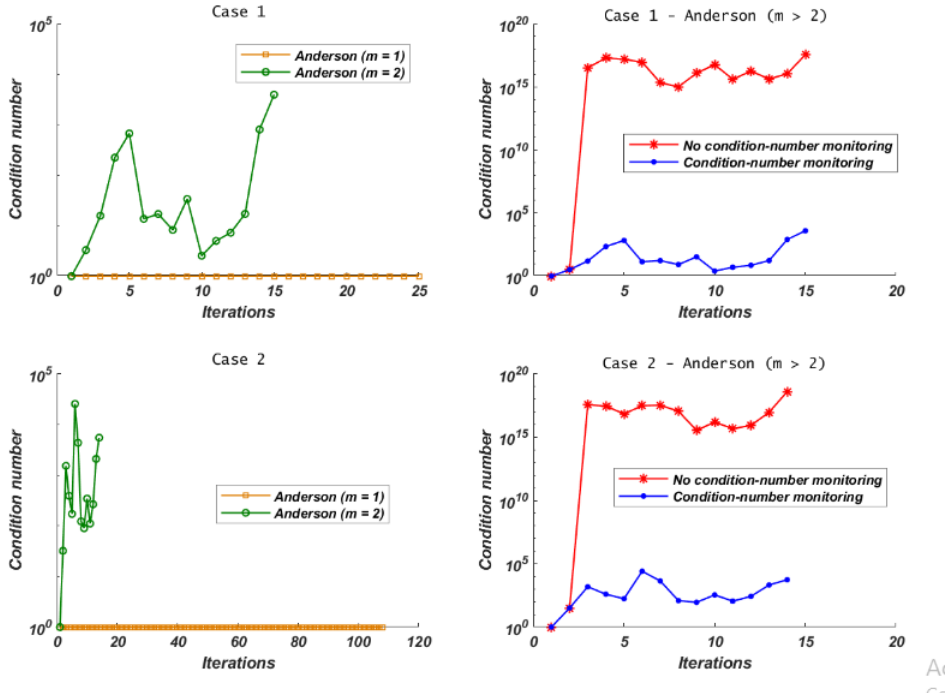

FIGURE 1.2 – Cas test d'acide gallique : Equilibre thermodynamique par la méthode AA - Nombre de conditionnement

mence après 12 ou 13 itérations pour les cas 1 et 2. Cependant, avec AA($m = 1$), la norme résiduelle diminuent lentement surtout dans le cas 2 où la convergence nécessite plus de 100 itérations. Plus précisément, pour le cas 2, si l'on prend $\kappa = 0.3$, AA($m = 1$) converge plus rapidement et la norme l_2 résiduelle diminue nécessitant 26 itérations, au lieu de 109, pour être inférieure à 10^{-10}. En revanche, les pentes "théoriques" de Newton sont tracées sur la Figure 1.3. Ces pentes montrent qu'une convergence d'ordre 2 est atteinte, ce qui confirme que la mé-

thode AA fonctionne d'une manière très efficace.

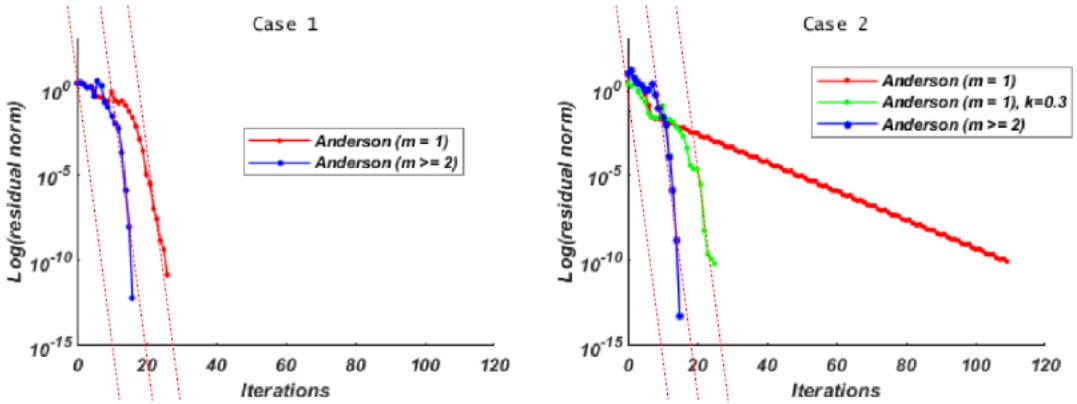

FIGURE 1.3 – Cas test d'acide gallique : Équilibre thermodynamique par la méthode AA - Norme résiduelle

4. **Temps de calcul CPU**
   Comme indiqué dans le tableau 1.3, la méthode AA (m) converge pour atteindre la solution de l'équilibre chimique du cas test d'acide gallique en un temps de calcul (CPU) très court et pour différentes valeurs de la profondeur maximale m ; $m \in \{1, 2, 3\}$.

TABLE 1.3 – Cas test d'acide guallique : Équilibre thermodynamique par la méthode AA - Temps de Calcul CPU (s)

	Temps CPU (s)
Anderson $(m = 1)$	1.11
Anderson $(m = 2)$	1.3
Anderson $(m = 3)$	1.13

Résultats des méthodes MPE et RRE Pour le cas test d'acide gallique, les vecteurs des logarithmes de concentration $\omega_1, \omega_2, \ldots$ sont générés par (1.80), en prenant différentes valeurs du paramètre κ, $\kappa \in \{0.1, 0.45, 0.5, 0.6\}$. Ces valeurs constituent des bons choix pour cette expérience, mais cela ne signifie pas que κ ne peut pas prendre une autre valeur. On applique les deux méthodes MPE et RRE en mode cyclique avec plusieurs choix pour le triplet $(Kmax, N_0, N)$:

- $(Kmax, N_0, N) = (10, 20, 10)$, $(20, 0, 10)$, $(10, 20, 0)$ pour $\kappa = 0.1$,
- $(Kmax, N_0, N) = (10, 10, 15)$,$(10, 10, 10)$ pour $\kappa = 0.45$,
- $(Kmax, N_0, N) = (10, 5, 15)$, $(20, 5, 15)$ pour $\kappa = 0.5$,
- $(Kmax, N_0, N) = (15, 15, 15)$ pour $\kappa = 0.6$.

1. **Taux de convergence**
   Les Figures 1.4 et 1.5 illustrent le comportement de la norme l_2 résiduelle, à l'aide d'une échelle logarithmique. La convergence est globalement linéaire, mais un petit mode marginalement instable est observé qui correspond à une oscillation semi-sinusoïdale quasi périodique des résidus. Avec $k = 0.1$ (cf. Figure 1.4), le premier choix des trois données est le meilleur car il donne la convergence la plus rapide de l'erreur résiduelle. En comparant les premier et le troisième choix des données $Kmax$, N_0 et N, on remarque que l'exécution de quelques itérations de base avant l'application de MPE ou RRE dans chaque cycle après le premier cycle permet

d'obtenir une convergence plus rapide ainsi qu'un comportement plus stable de la norme résiduelle, malgré cette légère oscillation. Un résultat similaire est observé dans la Figure 1.5 avec $\kappa = 0,45$, $\kappa = 0,5$ et $\kappa = 0.6$. Ces trois dernières valeurs de κ donnent encore une convergence rapide représentée par une diminution quasi stable de la norme résiduelle.

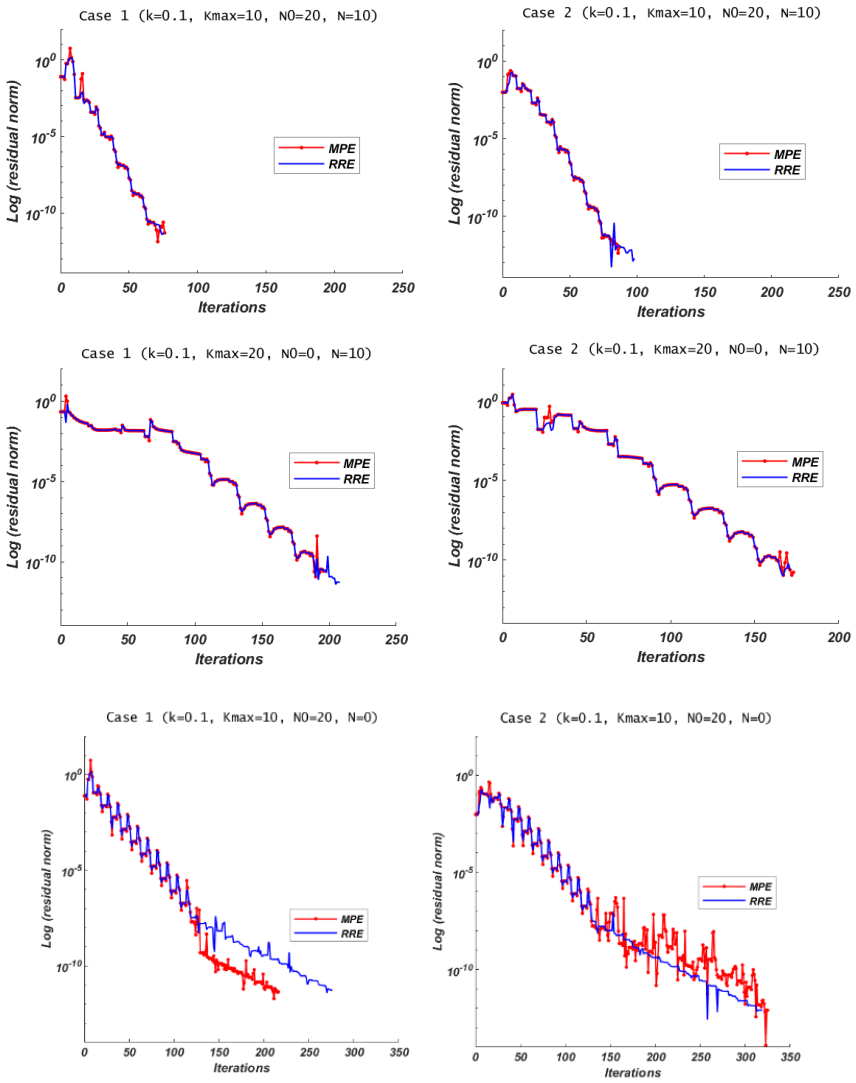

FIGURE 1.4 – Cas test d'acide gallique : Équilibre thermodynamique par les méthodes MPE et RRE redémarrées avec $\kappa = 0.1$ - Norme résiduelle

2. **Temps de calcul CPU**
   Pour l'équilibre chimique du cas test d'acide gallique, cas 1 et 2, le temps de calcul nécessaire pour effectuer le nombres totales d'itérations avec les méthodes redémarrées MPE et RRE est très court ne dépassant pas une seconde (cf. tableau 1.4 pour le cas 1).

À partir de ces observations, on constate que, lors de la résolution du système d'équilibre chimique pour le cas test d'acide gallique, la méthode d'Anderson semble être plus stable que les méthodes MPE et RRE, et nécessite le moins d'itérations. Ceci est dû à la stratégie de redémarrage appliquée aux méthodes MPE et RRE, alors que la méthode AA ne nécessite pas une telle stratégie.

1.3.3.2 Cas test 1D "Easy" de Benchmark MoMas

Présentation du test Le Benchmark MoMas a été conçu pour comparer les méthodes numériques pour le modèle de transport réactif en 1D et 2D. Différentes méthodes de couplage ont été utilisées

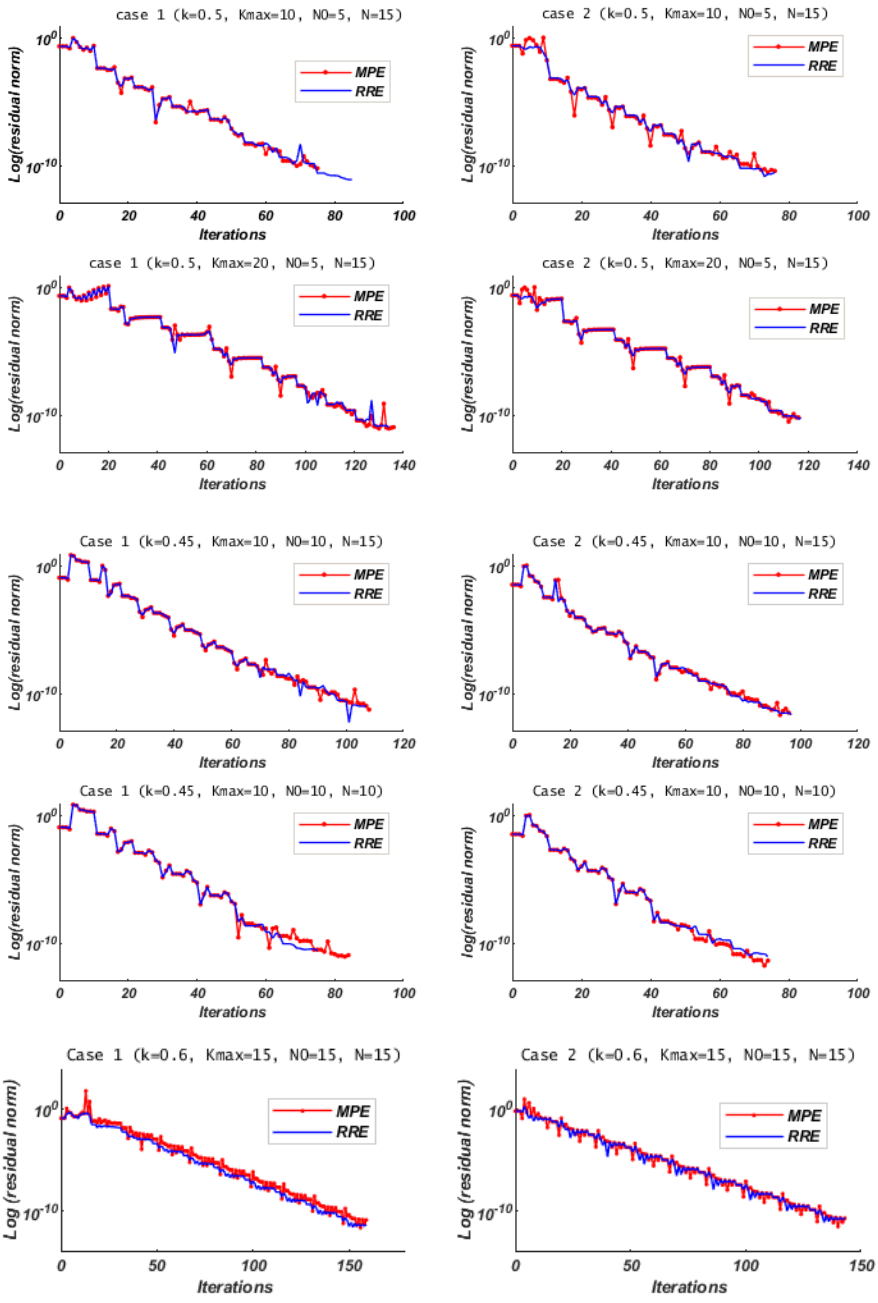

FIGURE 1.5 – Cas test d'acide gallique : Équilibre thermodynamique par les méthodes MPE et RRE redémarrées avec $\kappa = 0.45$, $\kappa = 0.5$ et $\kappa = 0.6$ - Norme résiduelle

pour résoudre ce critère. La définition est publiée dans [40] et les résultats des participants sont comparés dans l'article de synthèse [39]. Le Benchmark se compose de trois cas tests de difficulté progressive nommés "Easy", "Medium" et "Hard". Chaque cas test hérite du modèle chimique précédent avec une difficulté en plus. Ce test présente un grand challenge en dépit du fait que les cœfficients chimiques ne soient pas tirés d'un système chimique réel.
Dans ce travail, on ne s'intéresse qu'à la résolution de l'équilibre chimique du cas test "Easy", réalisé sur une géométrie 1D. Pour ce cas, le système chimique est composé de $n_e = 12$ espèces chimiques réparties comme suit :

- $n_{pm} = 4$ espèces primaires mobiles : X_1, X_2, X_3, X_4 ;
- $n_{pf} = 1$ espèce primaire fixe : S ;
- $n_{sm} = 5$ espèces secondaires mobiles : C_1, C_2, C_3, C_4,C_5 ;

TABLE 1.4 – Temps de calcul CPU (s) de l'équilibre thermodynamique pour le test de l'acide gallique (cas 1) par les méthodes MPE et RRE redémarrées

	$(Kmax, N_0, N)$	**MPE**	**RRE**
$\kappa = 0.1$	(10,20,10)	0.6406	0.8594
	(20,0,10)	0.6719	0.4688
	(10,20,0)	0.8594	0.3906
$\kappa = 0.45$	(10,10,10)	0.2656	0.2031
	(10,10,15)	0.4531	0.2969
$\kappa = 0.5$	(10,5,15)	0.2656	0.2656
	(20,5,15)	0.3906	0.5

- $n_{sf} = 2$ espèces secondaires fixes : CS_1, CS_2.

Le domaine $1D$ est considéré comme un tube de longueur $L = 2.1$, il est hétérogène et composé de deux sous-domaines A et B. Le sous domaine A est plus perméable avec une faible porosité et une faible réactivité, tandis que le sous domaine B est moins perméable avec de plus grandes porosité et réactivité.

Afin d'être proche des cas réalistes, les conditions initiales et aux limites ne sont pas exprimées pour les variables fondamentales, c'est-à-dire les concentrations des composants. En effet, l'analyse chimique peut fournir assez facilement une mesure de la concentration totale ou de la concentration totale dissoute pour chaque composant. Une injection est faite sur le côté gauche du domaine suivie d'un lessivage (*'leaching'*) sur le même côté. La période d'injection dure 5000 s alors que la période de lessivage durent au moins 1000 s. Cette dernière peut être rallongée, si nécessaire, pour atteindre la condition suivante : à la fin de la période de lessivage, 99,9% des espèces injectées (X_1, X_3 and S) sont déplacées du domaine.
La transition entre ces deux périodes se fait en imposant la condition aux limites suivante :

$$T_j(x = 0, t) = T_j^{inj} \quad \text{pour} \quad t < 5000\ s \quad \text{et} \quad T_j(x = 0, t) = T_j^{less} \quad \text{pour} \quad t > 5000\ s$$

Les espèces chimiques interagissent dans le cas "1D Easy" à travers $n_r = 7$ réactions à l'équilibre. Le domaine est initialement en équilibre avec l'espèce S en présence des espèces mobiles X_2 et X_4. L'espèce X_1 joue le rôle d'un traceur. Durant la période d'injection, le composant X_4 est lessivée du domaine, les espèces X_2 et X_3 interagissent avec la surface S et X_4 encore présentes. Durant la période du lessivage, les espèces X_1 et X_3 sont lessivées du domaine, les espèces X_2 et X_4 interagissent avec la surface S et X_3 encore présentes.

Dans ce cas test, une constante de réaction à l'équilibre est de l'ordre 10^{35}, ce qui fait que le système est très raide et présente déjà un grand challenge. De plus, les coefficients stœchiométriques sont assez grands, ce qui permet de tester la robustesse de notre implémentation face à une telle complexité. Le tableau de Morel 1.5 illustre les cœfficients stœchiométriques pour les lois d'action de masse et de conservation.

Dans ce qui suit, on présente tout d'abord les résultats de l'équilibre chimique dans les deux sous-domaines A et B en supposant, par exemple, que le vecteur des concentrations initiales des espèces

TABLE 1.5 – Tableau des équilibres pour le cas test "Easy" de Benchmark MoMas

Espèces	X_1	X_2	X_3	X_4	S	**Constante d'équilibre K**
C_1	0	-1	0	0	0	10^{-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	10^{35}
CS_1	0	3	1	0	1	10^{6}
CS_2	0	-3	0	1	2	10^{-1}
Concentration totale T	T_1	T_2	T_3	T_4	T_S	
Conditions initiales						
Sous-domaine A	0	-2	0	2	1	
Sous-domaine B	0	-2	0	2	10	
Condition aux limites						
Injection $t \in [0, 5000]$	0.3	0.3	0.3	0	0	
Lessivage $t \in [5000, ...]$	0	-2	0	2	0	

primaires dans chacun de ces deux sous-domaines est donné par :

$$\begin{aligned}\mathcal{X}_{A,B,0}\,(M) &= (X_{1,0}, X_{2,0}, X_{3,0}, X_{4.0}, S_0)^T \\ &= (0.3, 0.4, 10^{-11}, 0.21, 0.6)^T.\end{aligned}$$

Après avoir mis le domaine initialement à l'équilibre, on présente les résultats d'équilibre thermodynamique pour les périodes d'injection (dans A et B) et de lessivage.

Résultats de la méthode AA On étudie la convergence de la méthode d'Accélération d'Anderson pour toute valeur de la profondeur maximale $m > 0$ en considérant tout d'abord le problème d'équilibre chimique comme étant sans relaxation ($\kappa = 1$).

1. **Convergence vers la solution**

   (a) *Équilibre chimique dans les zones A et B*
   Pour l'équilibre chimique dans les deux sous-domaines A et B, les figures 1.6 et 1.7 montrent le profil des concentrations des espèces composants en fonction du nombre d'itérations effectué pour atteindre la convergence. Aucun phénomène d'oscillations compliquées n'est observé, ce qui veut dire que la convergence est atteinte sans difficulté. Notons que les concentrations des composants à l'équilibre thermodynamique dans les deux sous-domaines A et B sont définies respectivement par les deux vecteurs :

$$\begin{aligned}\mathcal{X}^*_A\,(M) &= (X^*_{1,A}, X^*_{2,A}, X^*_{3,A}, X^*_{4.A}, S^*_A)^T \\ &= (10^{-20}, 0.2597, 10^{-20}, 0.3495, 0.3907)^T \\ \mathcal{X}^*_B\,(M) &= (X^*_{1,B}, X^*_{2,B}, X^*_{3,B}, X^*_{4.B}, S^*_B)^T \\ &= (10^{-20}, 1.5116, 10^{-20}, 0.5756, 7.9128)^T\end{aligned}$$

   Ainsi, l'influence du sous-domaine le plus réactif B peut être vue à partir de la concentration la plus élevée de S atteinte à l'équilibre.

   (b) *Période d'injection*
   Deux cas sont étudiés : l'injection dans le sous-domaine A (à gauche) et l'injection dans

le sous-domaine B. Comme on a déjà mentionné, puisque le domaine est initialement en équilibre avec l'espèce S en présence des espèces mobiles X_2 et X_4 alors l'approximation initiale des concentrations des composants dans chaque sous-domaine est égale à la solution d'équilibre thermodynamique déjà obtenue, c'est à dire :

$$\mathcal{X}_{inj,0} = \begin{cases} \mathcal{X}_A^* & \text{injection dans le sous-domaine A} \\ \mathcal{X}_B^* & \text{injection dans le sous-domaine B.} \end{cases}$$

On remarque d'après les figures 1.8 et 1.9 qu'un petit nombre d'itérations est effectué afin d'atteindre la convergence de la méthode AA ($m > 0$) pour le problème d'équilibre chimique durant la période d'injection. La solution ainsi obtenue est donnée par le vecteur suivant :

$$\begin{aligned} \mathcal{X}_{inj}^*(\mathrm{M}) &= (\mathrm{X}_{1,inj}^*, \mathrm{X}_{2,inj}^*, \mathrm{X}_{3,inj}^*, \mathrm{X}_{4,inj}^*, \mathrm{S}_{inj}^*)^T \\ &= (0.3, 0.2416, 0.2416, 10^{-50}, 10^{-23})^T. \end{aligned}$$

Ce résultat vérifie qu'à l'équilibre, le composant X_4 est lessivé du domaine et les deux composants X_2 et X_3 restent présents pour interagir avec les surfaces S et X_4. De plus, la concentration de X_1 reste constante à 0,3 M parcequ'il joue le rôle d'un traceur.

(c) *Période de lessivage*
Après une période de 5000 s d'injection sur le côté gauche du domaine, une période de lessivage se déroule sur le même côté. Grâce à cette transition, on étudie l'équilibre thermodynamique pour la période de lessivage en considérant comme approximation initiale des concentrations des composants celles obtenues à l'équilibre chimique de la période d'injection, c'est à dire :

$$\mathcal{X}_{less,0} = \mathcal{X}_{inj}^*$$

En se référant à la figure 1.10, pour laquelle on considère le problème d'équilibre chimique comme étant sans relaxation, AA ($m = 2$) et AA ($m = 3$) convergent sans difficultés nécessitant 61 et 39 itérations respectivement. Cependant, après plusieurs essaies numériques, on constate que AA ($m = 1$) et AA ($m \geq 4$) ne convergent pas ou présentent parfois des difficultés de convergence. Cela est peut-être dû au choix des concentrations initiales. Ces difficultés sont résolues en ajoutant au problème un terme de relaxation $\kappa \neq 1$ choisi arbitrairement jusqu'à ce qu'il soit bien adapté pour atteindre la convergence. On prend respectivement $\kappa = 0.29,\ \kappa = 0.55,\ \kappa = 0.46$ et $\kappa = 0.42$ pour AA($m = 1$), AA($m = 4$), AA($m = 5$) et AA($m > 5$). Ainsi, à l'équilibre chimique de la période de lessivage, les concentrations des composants sont données par le vecteur solution suivant :

$$\begin{aligned} \mathcal{X}_{less}^*(\mathrm{M}) &= (\mathrm{X}_{1,leach}^*, \mathrm{X}_{2,less}^*, \mathrm{X}_{3,less}^*, \mathrm{X}_{4,less}^*, \mathrm{S}_{less}^*)^T \\ &= (10^{-20}, 5, 7735.10^{-7}, 7.223.10^{-27}, 1.1547.10^{-6}, 10^{-20})^T \end{aligned}$$

2. **Conditionnement de $\mathcal{F}_k$**
Puisqu'une stratégie de surveillance du nombre de conditionnement est utilisée lors de l'Accélération d'Anderson, il n'est pas nécessaire de s'inquiéter du fait que la matrice $\mathcal{F}_k$ deviendrait mal conditionnée. Cependant ce dernier point est surveillé pour garantir la stabilité et la robustesse de la méthode AA. Appliquée au système chimique du cas test 1D "Easy" de benchmark MoMas, l'effet de cette stratégie est très clair pour maintenir le conditionnement de la matrice $\mathcal{F}_k$ inférieur à 10^{10} :

(a) *Équilibre chimique dans les zones A et B*
On remarque que pour $m \geq 5$, $cond\ (\mathcal{F}_k)$ devient inférieur à 10^{10} pour $k \geq 6$ après avoir été supérieur à 10^{15} (cf. figures 1.11 et 1.12).

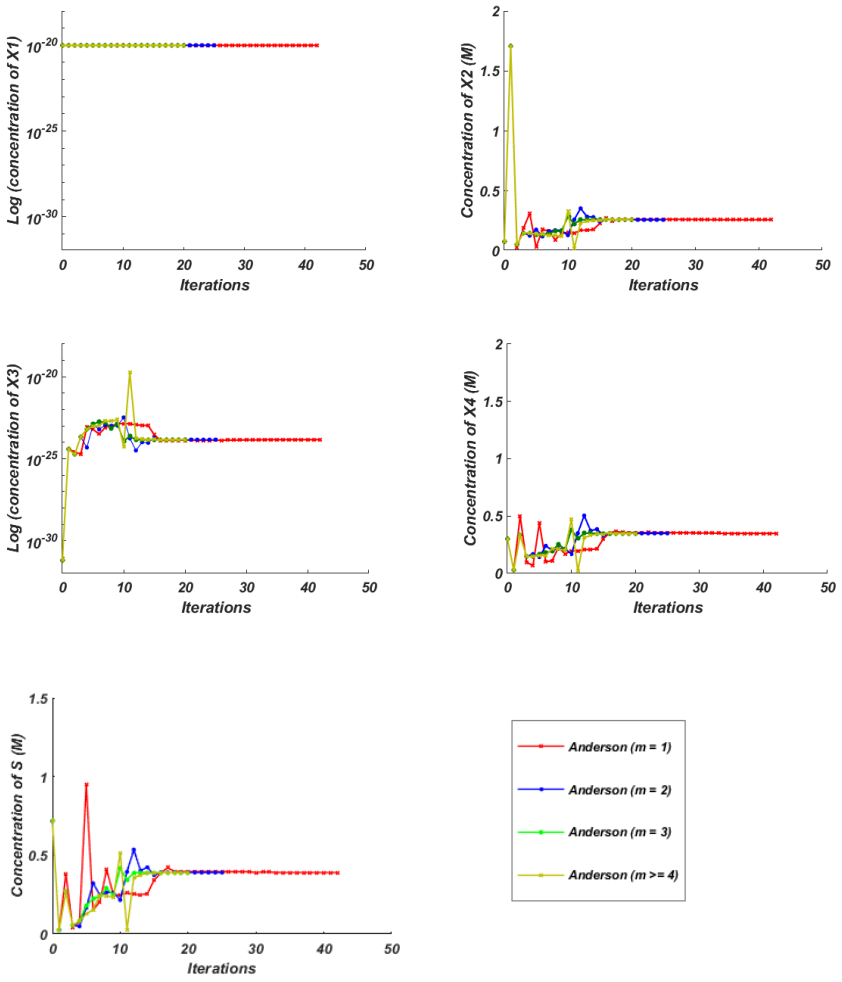

FIGURE 1.6 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique dans le sous-domaine A par la méthode AA

(b) *Période d'injection*
On note que lorsque $m = 4$ et $m \geq 5$, cette stratégie de contrôle est appliquée à la matrice $\mathcal{F}_k$ respectivement pour $k \geq 13$ (resp. $k \geq 9$) et $k \geq 6$ (resp. $k \geq 5$) si l'injection est dans le sous-domaine A (resp. B) (cf. figures 1.13 et 1.14).

(c) *Période de lessivage*
Avec AA ($m \geq 5$), $cond\,(\mathcal{F}_k)$ est surveillé durant plusieurs étapes d'itération k (cf. figure 1.15) :

$$\begin{cases} k \in \{2, 11, 12, 13, 15, 17, 18, 22\} & \text{si} \quad m = 5 \\ 6 \leq k \leq 16, 18 \leq k \leq 20 & \text{si} \quad m > 5 \end{cases}$$

3. **Taux de convergence**
La Figure 1.16 illustre le taux de convergence avec la méthode d'Anderson ($m = 1, 2, 3, 4, 5$) pour tous les cas du test "1D Easy" du benchmark MoMas, en faisant apparaître les pentes "théoriques" approximatives de Newton.

(a) *Équilibre chimique dans les zones A et B*
Pour les deux systèmes chimiques qui conviennent, AA ($m = 1$) nécessite environ le double du nombre d'itérations effectué par AA ($m = 3, 4$) pour atteindre le niveau résiduel prescrit à l'équilibre.

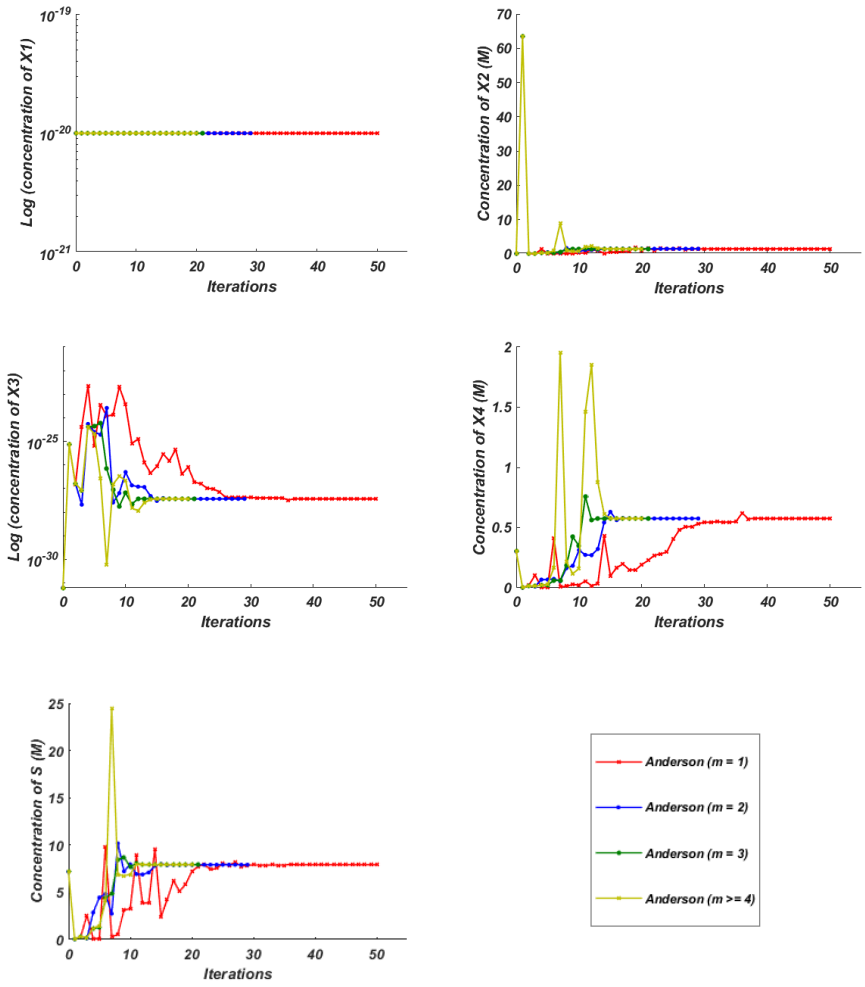

FIGURE 1.7 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique dans le sous-domaine B par la méthode AA

(b) *Période d'injection*
Durant l'injection dans la zone A, on remarque que la norme résiduelle diminue plus rapidement avec AA ($m = 2$). D'autre part, dans la zone B, la norme résiduelle diminue avec un taux de convergence raisonnable pour $m = 1, 2, 3$ et 4.

(c) *Période de lessivage*
Pour le système chimique décrivant cette période, on note que AA($m = 4$) converge le plus rapidement, AA ($m = 2, 3$) montre des problèmes d'oscillations résiduelles après quelques itérations, avant que la convergence ne reprenne son chemin stable et AA ($m = 1$) montre un retard décrit par un résidu constant entre les itérations 39 et 60.

Compte tenu de ces résultats, les pentes prouvent la convergence d'ordre 2 de la méthode AA et montrent de nouveau son efficacité.

4. **Temps de calcul CPU**
Nous tabulons le temps d'exécution CPU requis par la méthode AA (m) pour $m \in \{1, 2, 3, 4\}$ afin de résoudre l'équilibre thermodynamique dans chaque sous-domaine et pendant les deux périodes d'injection et de lessivage. La rapidité de cette méthode se présente clairement par un très court temps de calcul ne dépassant pas 2 secondes (cf. tableau 1.6).

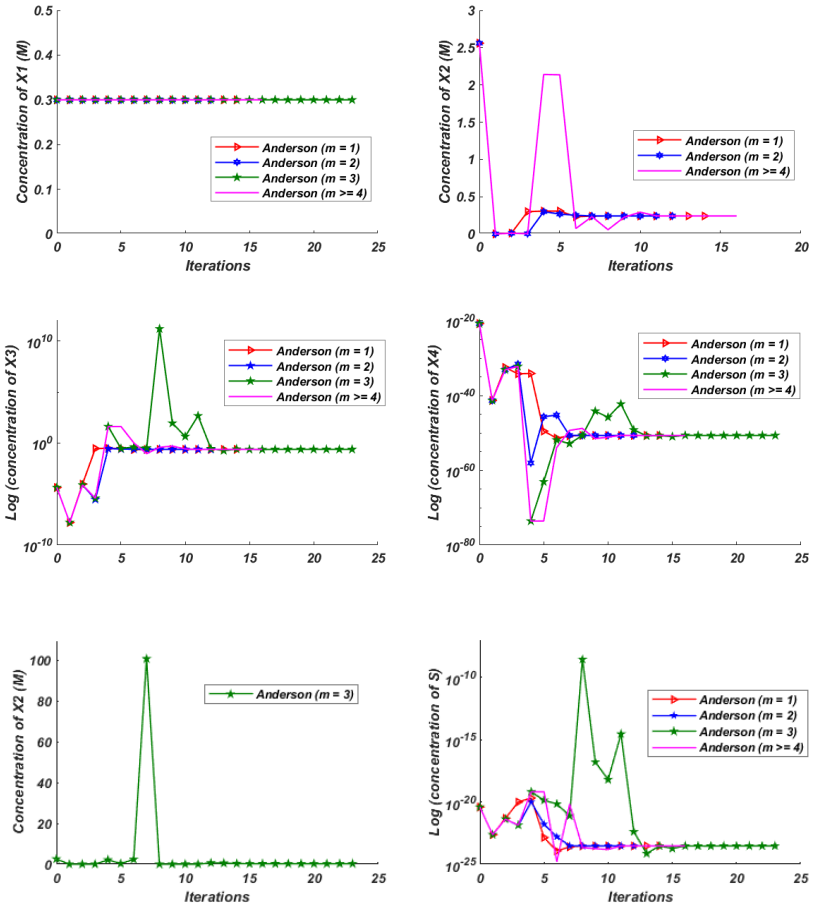

FIGURE 1.8 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période d'injection dans A par la méthode AA

Résultats des méthodes MPE et RRE On applique les méthodes MPE et RRE en mode cyclique au système non linéaire d'équilibre thermodynamique du cas test "1D Easy" de benchmark MoMas en générant les vecteurs des logarithmes de concentration $\omega_1, \omega_2, \ldots$ par (1.80), où le paramètre de relaxation κ varie selon les cas de ce cas test ; il est choisi arbitrairement pour assurer la convergence de la manière la plus efficace. Notre code de programme informatique est exécuté en supposant que le nombre maximum d'itérations effectuées dans chaque cycle est $Kmax = 10$ et une fois $kmax = 15$ (dans un cas du système d'injection).

1. **Taux de convergence**

   (a) *Équilibre chimique dans la zone A*

   Pour le système d'équilibre chimique dans le sous-domaine A, on prend $\kappa = 0.4$. Plusieurs choix du couple (N, N_0) sont considérés. La figure 1.17 montre l'évolution de la norme résiduelle non linéaire, en utilisant une échelle logarithmique pour les méthodes MPE et RRE redémarrées. Pour $(N, N_0) = (0, 20), (5, 20)$, ces deux méthodes convergent vers un état stationnaire. De plus, lorsque l'on effectue un certain nombre d'itérations avant que RRE ou MPE ne soit appliqué dans chaque cycle après le premier cycle, la norme résiduelle diminue plus rapidement et la convergence est atteinte en un nombre d'itérations plus réduit ; ceci est clair où l'on suppose que $(N, N_0) = (5, 20)$. Pour $(N, N_0) = (20, 0)$, on remarque que MPE converge plus rapidement que RRE, on observe néanmoins quelques perturbations décrites par une augmentation résiduelle entre les itérations 15 et 20. Enfin, avec $(N, N_0) = (10, 0), (5, 0)$, RRE et MPE semblent avoir des performances similaires et convergent avec stabilité.

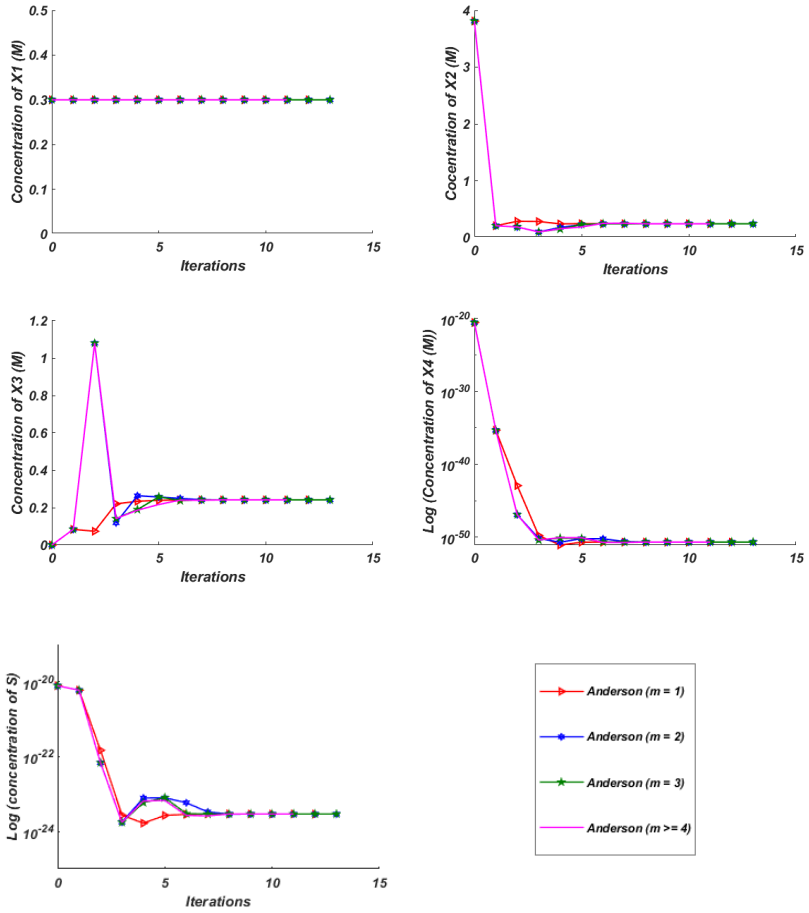

FIGURE 1.9 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période d'injection dans B par la méthode AA

(b) *Équilibre chimique dans la zone B*
On considère le système d'équilibre dans le sous-domaine B avec $k = 0.3$. La figure 1.18 illustre l'historique résiduel obtenu avec $(N, N_0) = (15, 0), (0, 80)$ et montre que, pour $(N, N_0) = (0, 80)$, il faut près de 300 itérations pour atteindre le niveau de convergence prescrit. Cependant, le meilleur résultat est obtenu avec $(N, N_0) = (15, 0)$ où les algorithmes des méthodes RRE et MPE prennent environ 90 et 80 itérations respectivement pour atteindre le niveau résiduel prescrit. Notons que les expériences numériques qu'on a faites pour cet exemple montrent que si l'on prend $N = 0$, il faut effectuer un grand nombre N_0 d'itérations de base pour atteindre la convergence.

(c) *Période d'injection*
Pour le le système d'équilibre chimique décrivant la période d'injection dans chaque sous-domaine, deux choix du paramètre κ sont considérés : $\kappa = 0.2$ et $\kappa = 1$. Pour les mêmes valeurs de $Kmax$, N et N_0, MPE et RRE semblent fonctionner de manière très similaire et atteignent la même précision dans les deux cas d'injection (injections dans A et B) (cf. figures 1.19 et 1.21), sauf pour le choix de $N_0 = 25, N = 0$ dans la Figure 1.21. Ce choix présente une différence entre les résultats de deux méthodes : le résultat de la méthode MPE diminue par rapport à celui de la méthode RRE à partir du cinquième cycle et reprend de nouveau son chemin stable et constant. Par contre, en prenant le paramètre de relaxation κ différent de 1, les deux méthodes donnent de nouveau presque les mêmes résultats (voir Figure 1.19). Sinon, en gardant $\kappa = 1$, ceci est encore réalisé en prenant le paramètre N non nul ($N = 2$) (cf. Figure 1.20). Par conséquent, les paramètres impliqués

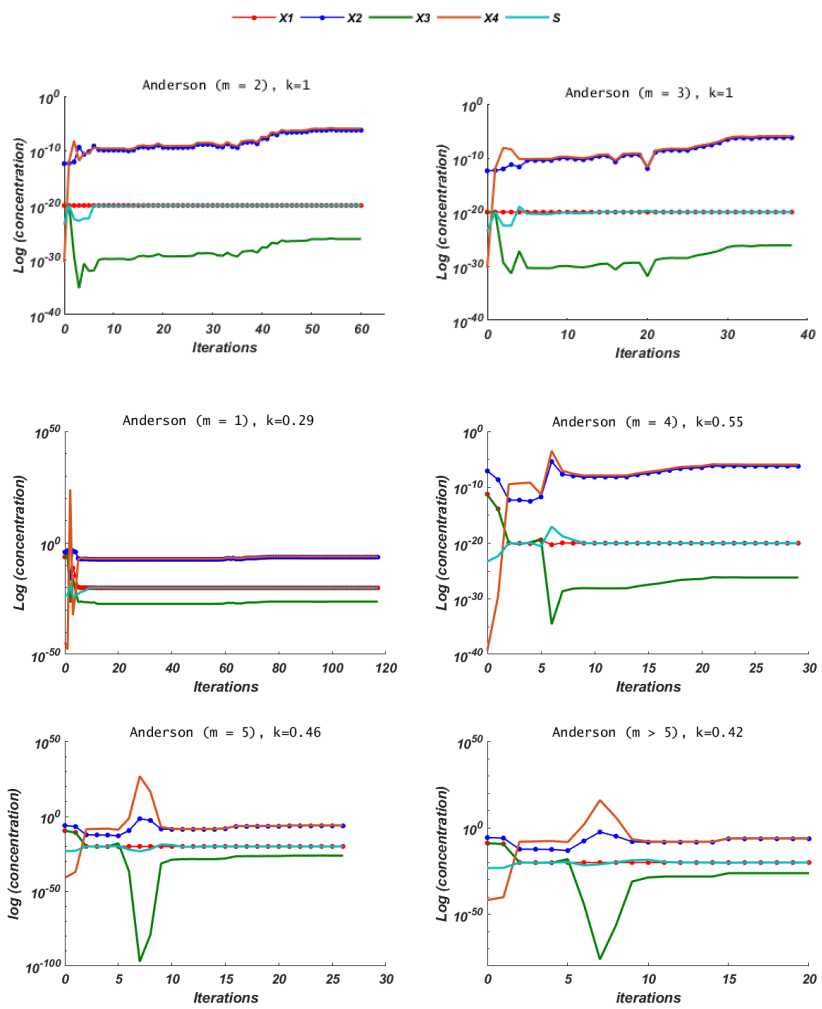

FIGURE 1.10 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période de lessivage par la méthode AA

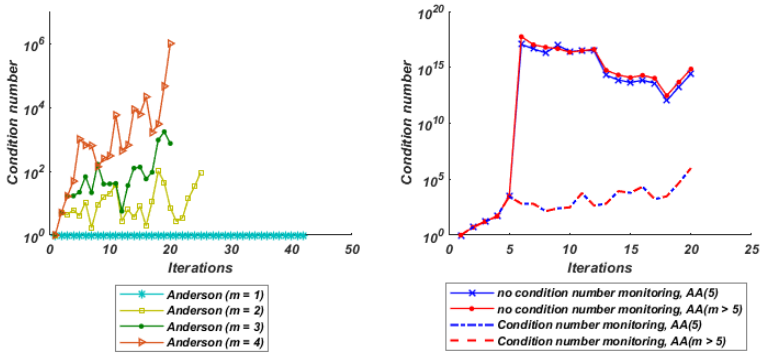

FIGURE 1.11 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique dans le sous-domaine A par la méthode AA - Nombre de conditionnement

dans l'implémentation numérique doivent être bien choisis afin que MPE et RRE donnent des résultats cohérents.
De plus, on note la diminution du nombre de cycles effectués pour une grande valeur de N, ce qui réduit le temps de calcul lorsque la surcharge temporelle du cyclage est prise en compte. Ce constat confirme l'efficacité de la stratégie du redémarrage. Notons que

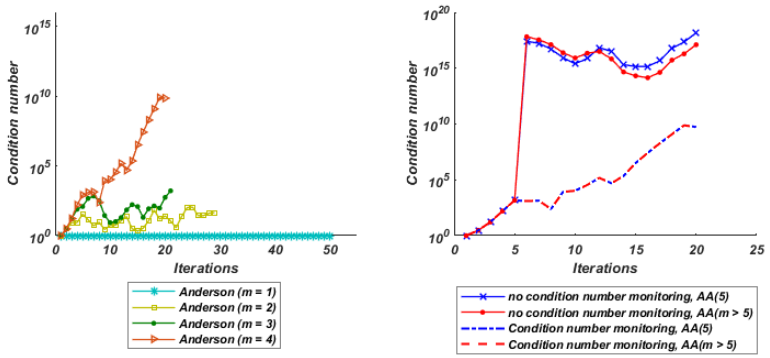

FIGURE 1.12 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique dans le sous-domaine B par la méthode AA - Nombre de conditionnement

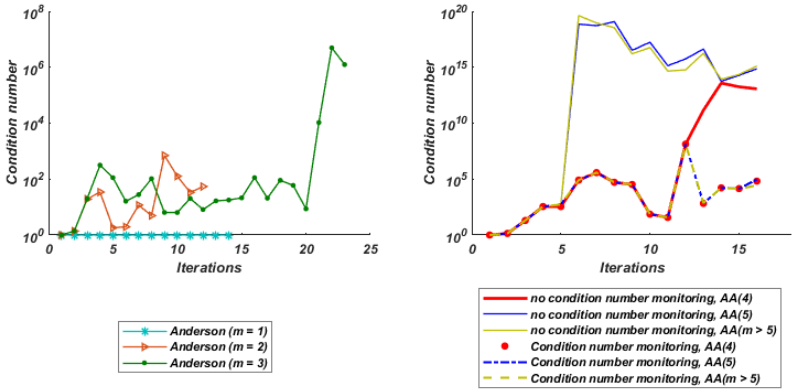

FIGURE 1.13 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période d'injection dans A par la méthode AA - Nombre de conditionnement

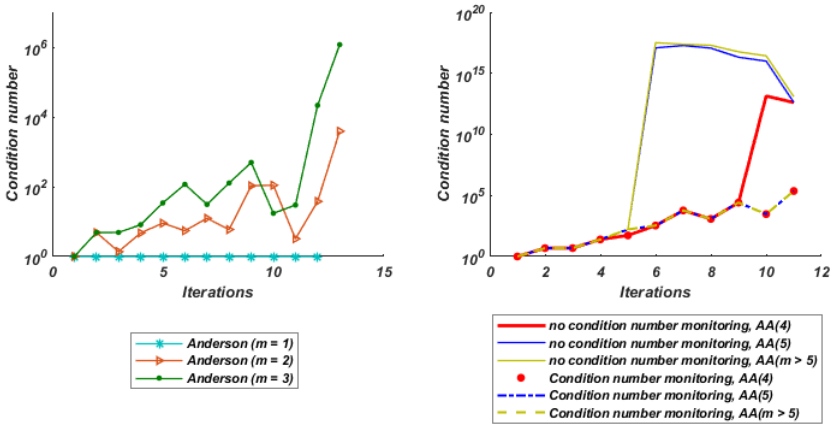

FIGURE 1.14 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période d'injection dans B par la méthode AA - Nombre de conditionnement

la convergence est beaucoup plus rapide pour $\kappa = 1$ que pour $k = 0,2$. Lorsque $\kappa = 1$, un très bon résultat est obtenu avec $(N, N_0) = (0,25), (0,18)$ où le niveau de convergence prescrit est atteint très rapidement dès la première itération, et encore où la norme résiduelle semble être une constante inférieure à la tolérance indiquée à ce niveau avec $(N, N_0) = (0,28), (0,24)$.

(d) *Période de lessivage*
On considère $k = 0,495$. Ce choix du paramètre κ est le meilleur pour atteindre la conver-

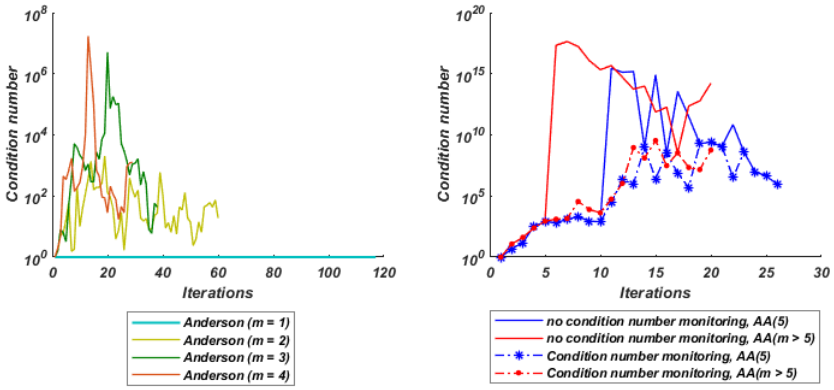

FIGURE 1.15 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période de lessivage par la méthode AA - Nombre de conditionnement

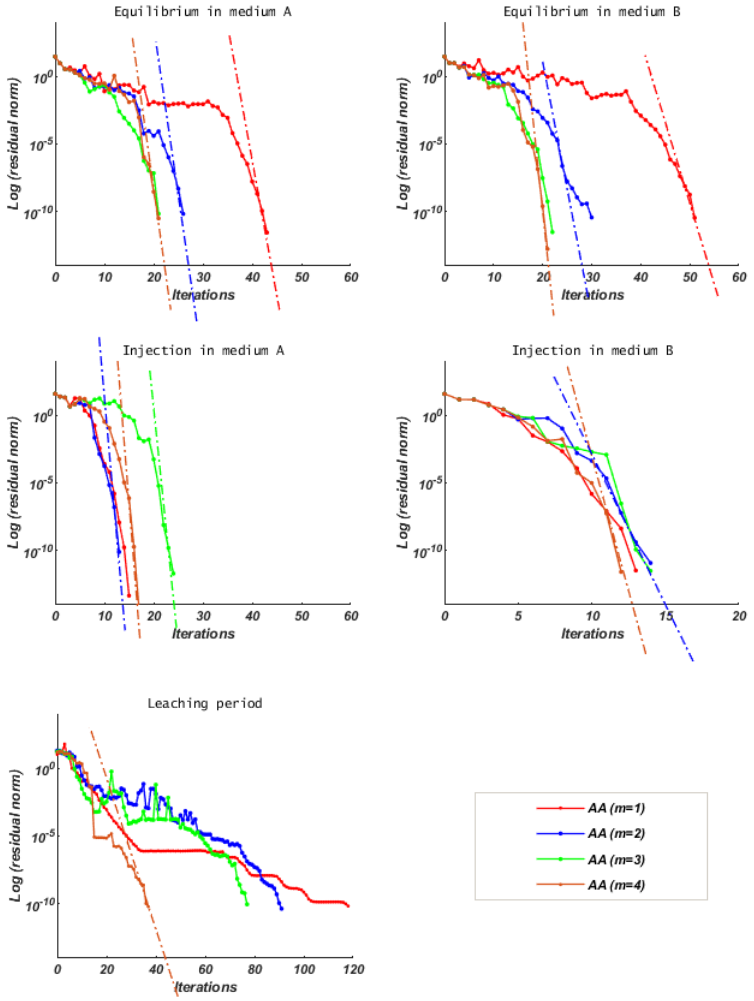

FIGURE 1.16 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique par la méthode AA - Norme résiduelle

gence même s'il entraîne des difficultés de convergence. On observe un mode instable présenté par des oscillations résiduelles (cf. figure 1.22). Pour $(N_0, N) = (0, 10)$, la convergence des méthodes RRE et MPE nécessitent les mêmes nombres d'itérations et de cycles, cependant pour $(N_0, N) = (10, 10)$, MPE converge plus rapidement que RRE contrairement au cas où $(N_0, N) = (5, 18)$. Cet exemple semble être un cas critique dans le sens

TABLE 1.6 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique par la méthode AA - Temps de calcul CPU (s)

	AA (m)	**Temps CPU (s)**
Sous-domaine A	$m = 1$	1.1875
	$m = 2$	1.0156
	$m = 3$	1.0781
	$m = 4$	1.2031
Sous-domaine B	$m = 1$	1.6719
	$m = 2$	1.3281
	$m = 3$	1.4063
	$m = 4$	1.6563
Période d'injection	$m = 1$	1.0156
	$m = 2$	1.0625
	$m = 3$	1.1094
	$m = 4$	1.9844
Période de lessivage	$m = 1$	1.3594
	$m = 2$	1.4063
	$m = 3$	1.2656
	$m = 4$	1.0313

qu'aucune règle de convergence ne peut être déduite en faisant varier les valeurs de N et N_0.

Par conséquent, pour résoudre le système d'équilibre chimique durant la période de lessivage, la méthode d'Accélération d'Anderson paraît plus efficace que les méthodes MPE et RRE redémarrées, en particulier pour une profondeur maximale $m = 4$ où la convergence est réalisée sans difficulté, avec une diminution résiduelle stable bien observée dans la figure 1.16.

2. **Temps de calcul CPU**
On résume tous les résultats de calcul du cas test "1D Easy" de benchmark MoMas dans le tableau 1.7. Pour tous les systèmes chimiques de ce cas test, le temps de calcul CPU requis pour effectuer le nombre total d'itérations $\mathbf{N_{It}}$ ainsi que celui des cycles $\mathbf{N_C}$ est très court ne dépassant pas 3 secondes, ce qui illustre l'efficacité et la robustesse des méthodes MPE et RRE, notamment en mode cyclique.

1.3.4 Comparaison avec les résultats des méthodes de types Newton-Raphson

Cas test de l'acide gallique Le calcul de l'équilibre thermodynamique du test de l'acide gallique présente des difficultés de convergence pour les méthodes de types Newton-Raphson [29, 22]. En suivant l'évolution du processus de recherche de la solution pour le cas 1 (cf. figure 2(a) de l'article [22]), on peut résumer les principales remarques suivantes :

- la méthode de Newton-Raphson ne permet pas d'obtenir la convergence dans le cas 1, pour lequel le processus de recherche est capturé par des oscillations ;
- la méthode du Simplex nécessite des temps de calcul importants, car la procédure de recherche tatônne longtemps loin de la solution ;
- avec la méthode de Newton-Raphson modifiée par *polishing factor*, on remarque la présence d'oscillations situées au voisinage proche de la solution ;

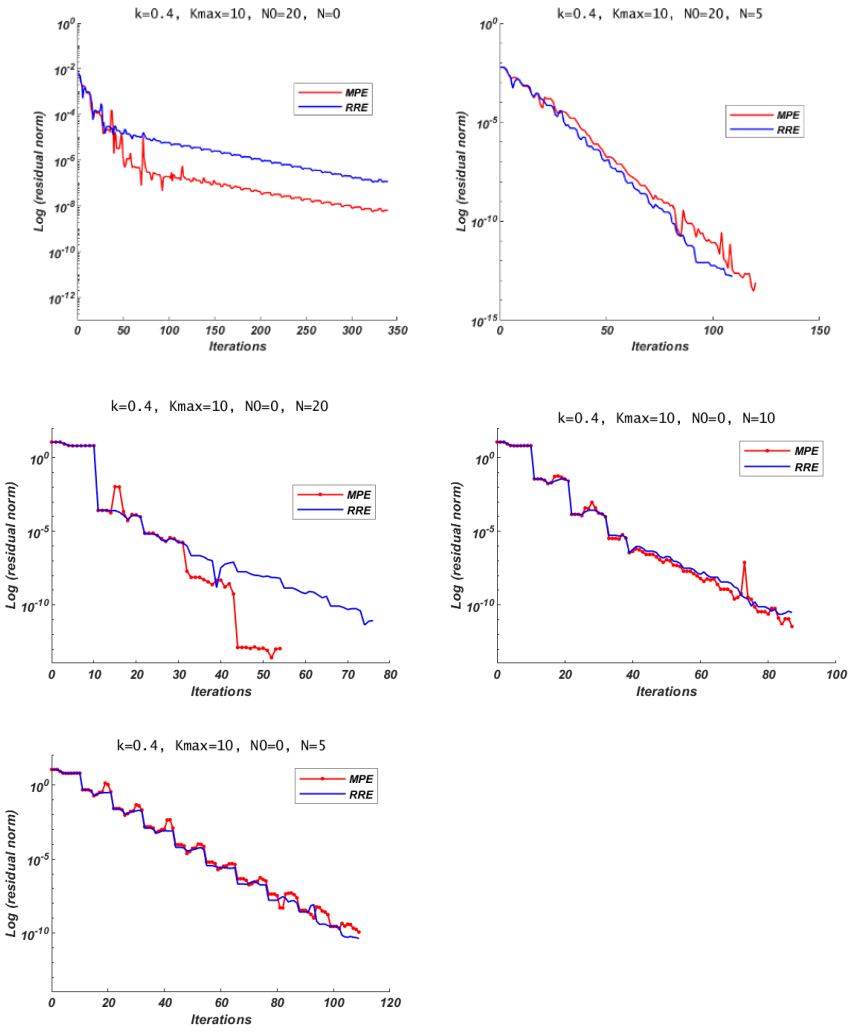

FIGURE 1.17 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique dans le sous-domaine A par les méthodes MPE et RRE redémarrées, avec $\kappa = 0.4$ - Norme résiduelle.

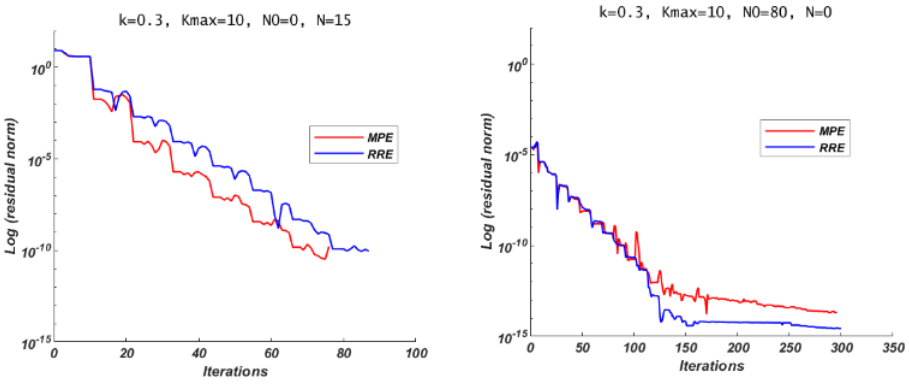

FIGURE 1.18 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique dans le sous-domaine B par les méthodes MPE et RRE redémarrées, avec $\kappa = 0.3$ - Norme résiduelle.

- la méthode de Newton-Raphson modifiée en imposant le respect du CAI (*chemical allowed interval*) montre des problèmes de convergence dont les oscillations responsables sont à l'intérieur du CAI ;

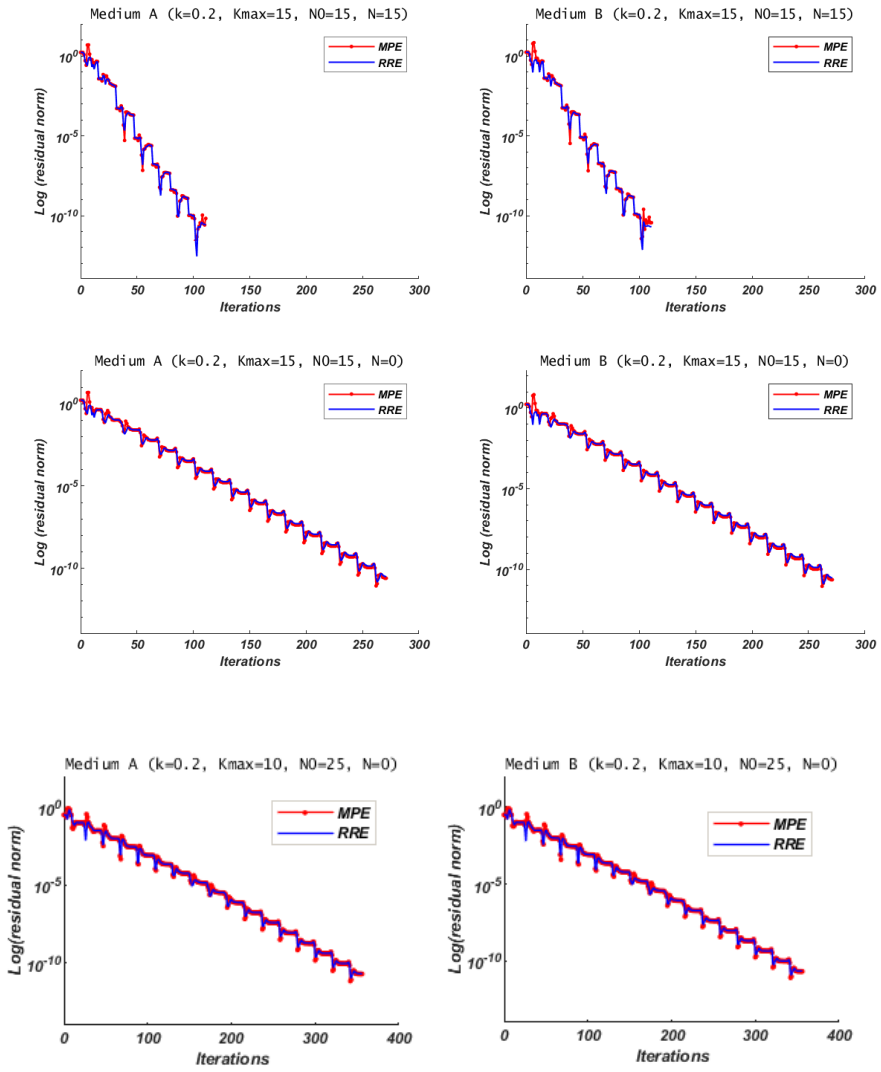

FIGURE 1.19 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période d'injection dans A et B par les méthodes MPE et RRE redémarrées, avec $\kappa = 0.2$ - Norme résiduelle

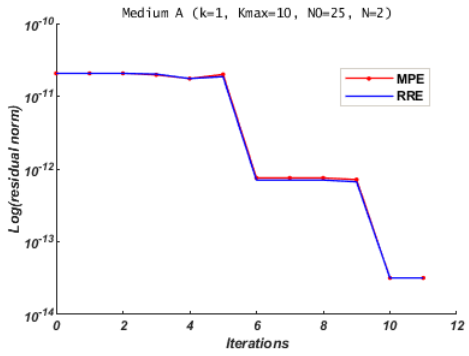

FIGURE 1.20 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période d'injection dans A et B par les méthodes MPE et RRE redémarrées, avec $\kappa = 1, N_0 = 25, N = 2, Kmax = 10$ - Residual norm curve

- avec la méthode des fractions continues positives, on voit que la convergence est fine et demande un temps de calcul long ;

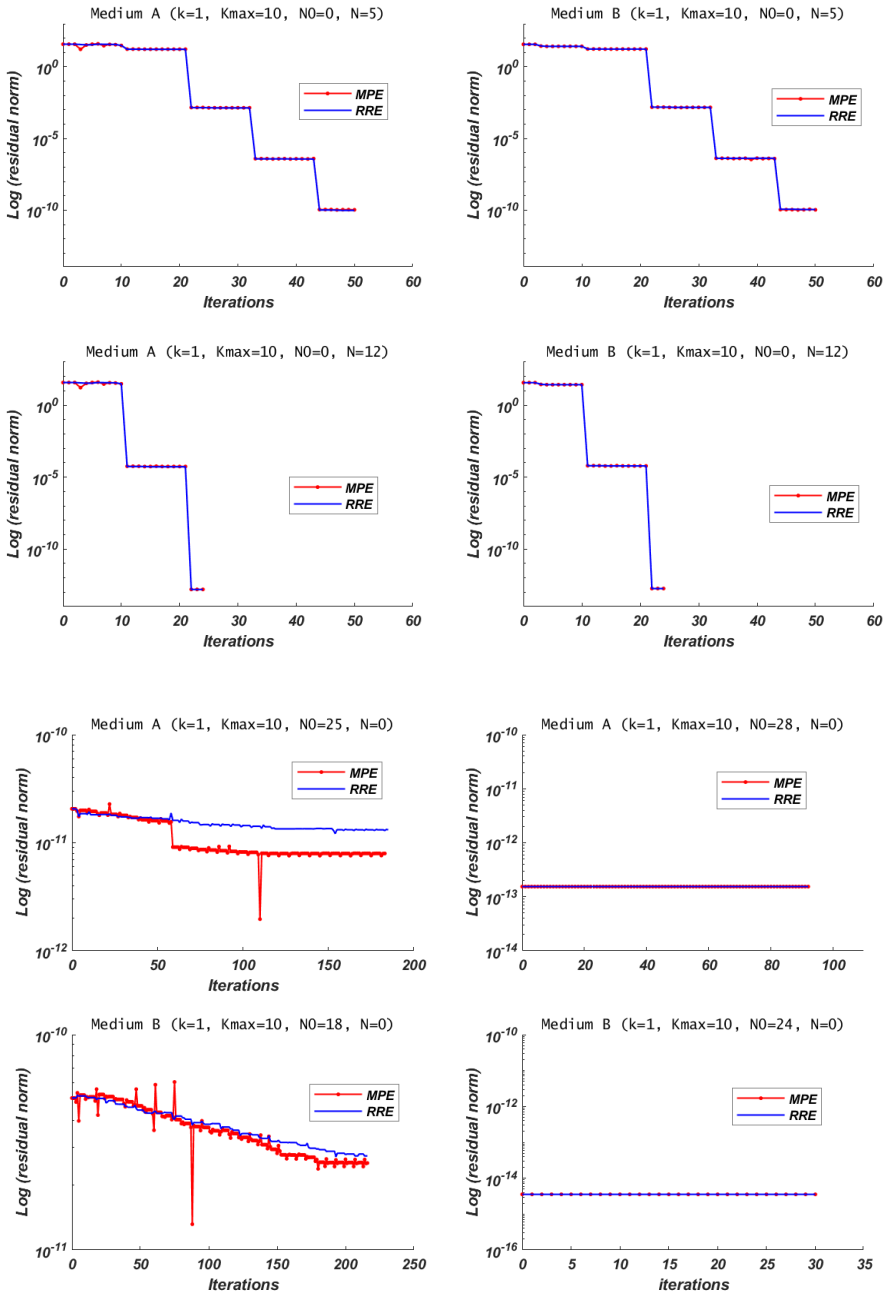

FIGURE 1.21 – Cas test "1D Easy" de benchmark MoMas : Équilibre thermodynamique de la période d'injection dans A et B par les méthodes MPE et RRE redémarrées, avec $\kappa = 1$ - Norme résiduelle

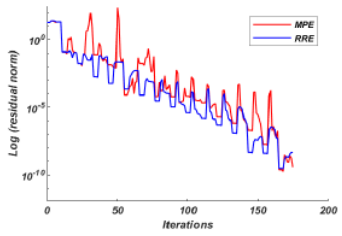

FIGURE 1.22 – Cas test "1D Easy" de benchmark MoMas : Équilibre chimique de la période de lessivage par les méthodes MPE et RRE redémarrées - Norme résiduelle

TABLE 1.7 – Cas test "1D Easy" de benchmark MoMas : Équilibre chimique par les méthodes MPE et RRE redémarrées - Temps de calcul CPU (s)

	$(\mathbf{Kmax}, \mathbf{N}, \mathbf{N_0})$	$(\mathbf{N_{It}}, \mathbf{N_C})$		**CPU (s)**	
		MPE	**RRE**	**MPE**	**RRE**
Zone A, $\kappa = 0.4$	$(10, 0, 20)$	(330,30)	(330,30)	1.0469	1.9531
	$(10, 5, 20)$	(121,11)	(110,10)	1.3906	0.7969
	$(10, 20, 0)$	(55,5)	(77,7)	0.9063	0.7188
	$(10, 10, 0)$	(88,8)	(88,8)	0.9844	1.3594
	$(10, 5, 0)$	(110,10)	(110,10)	1.0625	0.8750
Zone B, $\kappa = 0.3$	$(10, 15, 0)$	(77,7)	(88,8)	1.5469	0.7656
	$(10, 0, 80)$	(291,30)	(295,30)	1.7656	1.3438
Injection, $\kappa = 0.2$	$(15, 15, 15)$	(112,7)	(112,7)	1.2188	0.7969
	$(15, 0, 15)$	(256,17)	(256,17)	1.5781	1.4688
Injection, $\kappa = 1$	$(10, 5, 0)$	(51,5)	(51,5)	0.8750	0.8594
	$(10, 12, 0)$	(25,3)	(25,3)	0.5625	0.7813
	$(10, 0, 25)$ **dans A**	(178,30)	(180,30)	1.2031	1.1250
	$(10, 0, 28)$ **dans A**	(90,30)	(90,30)	1.2031	0.9844
	$(10, 0, 18)$ **dans B**	(210,30)	(210,30)	1.3594	1.3281
	$(10, 0, 24)$ **dans B**	(30,30)	(30,30)	0.8594	0.6875
Lessivage, $\kappa = 0.495$	$(10, 10, 0)$	(176,16)	(176,16)	1.2813	1.3438
	$(10, 10, 10)$	(154,14)	(231,21)	1.4688	2.2656
	$(10, 18, 5)$	(154,14)	(110,10)	2.1250	1.5313

Conclusion : Tous les résultats obtenus par des méthodes de type Newton-Raphson sont comparables à ceux qu'on a obtenus avec les méthodes AA, MPE et RRE, en terme d'oscillations et de temps de calcul. En effet, le processus de recherche de la solution dans les cas 1 et 2 pour la méthode d'AA ne présente aucun phénomène d'oscillations compliquées (cf. figure 1.1) ; il permet d'obtenir très rapidement une approximation grossière de la solution, en un petit nombre d'itérations et en un petit temps de calcul pour toute valeur strictement positive de m (cf. tableau 1.3). De même, pour les deux méthodes d'extrapolation polynomiales, MPE et RRE, on a déjà souligné l'obtention rapide de la solution sans aucune difficulté de convergence.

Cas test "1D Easy" de benchmark MoMas Pour ce cas test, on compare nos résultats avec ceux obtenus par des codes de transport réactif qui ont participé à la réalisation du benchmark [19, 39, 41] (HYTEC, SPECY, MIN3P, GDAE et *Hoffmann et al*). Le code HYTEC est appliqué au benchmark easy MoMas en tant que tel, sans qu'aucune modification ne soit apportée pour fonctionner plus rapidement ou pour améliorer la convergence, en prenant la précision de la résolution des équations chimiques (Newton- Raphson, code CHESS) égale à 10^{-8}. Loin des résultats concernant le problème de transport, les résultats de spéciation chimique qu'on a obtenu avec Anderson Acceleration, MPE et RRE sont en bon accord avec ceux obtenus par le code CHESS (cf. Table 4 de l'article [19]) (qui utilise la méthode Newton-Raphson) et ceci est clair d'après le tableau 1.8 (toutes les concentrations inférieures à 10^{-20} sont notées par "-" dans la partie droite du tableau 1.8).
Pour les quatre autres codes de transport réactif (SPECY, MIN3P, GDAE et Hoffmann et al), une solution de référence est donnée par le calcul du code SPECY [41] et une comparaison des résultats est présentée dans [39]. Bien que les résultats d'équilibre chimique ne soient pas présentés indépendamment lors du transport réactif, on peut donner une comparaison avec nos résultats de spéciation

chimique. Par exemple, comme les simulations des codes de transport réactif [41], le résultat du comportement de la concentration C_2 qu'on a obtenu durant les périodes d'injection et de lessivage avec les méthodes AA, MPE et RRE (cf. figure 1.23) reproduisent de manière constante l'augmentation et la diminution du front de la concentration en C_2 (cf. figure 7 de l'article [39]). D'autre part, la concentration (la plus élevée) du composant fixe S dans le sous-domaine B à l'équilibre montre clairement l'influence de la grande réactivité de cette zone ($S_B^* = 7.9128$ M), ce qui est en accord avec les résultats des codes de transport réactifs, dans le cas advectif (cf. figure 5 de l'article [39]), où, au temps 10, cette élévation de concentration se présente au centre du domaine, dans B, pour $1 \leq x \leq 1.1$ (x désigne la position dans l'espace). De plus, on remarque que pour $0 \leq x \leq 1$ et $1.1 \leq x \leq 2.1$, c'est à dire dans A, la concentration de S converge vers la même solution ($S_A^* = 0.39074$ M) qu'on a obtenue pour l'équilibre chimique dans A.

TABLE 1.8 – Comparaison des résultats de la spéciation chimique dans les zones initiales

	AA - MPE - RRE				**Newton-Raphson (code CHESS) [19]**			
	zone A	zone B	Injection	Lessivage	zone A	zone B	Injection	Lessivage
espèces								
X_1	1e-20	1e-20	0.3	1e-20	-	-	0.3	-
X_2	0.2597	1.5116	0.2416	5.7734e-07	0.25972	1.5116	0.24162	5.7735e-07
X_3	1.4604e-24	3.6593e-28	0.2416	7.2169e-27	-	-	0.24162	-
X_4	0.3495	0.5756	2.0800e-51	1.1547e-06	0.34954	0.57561	-	1.1547e-06
C_1	3.8503e-12	6.6157e-13	4.1387e-12	1.7321e-06	3.8503e-12	6.6157e-13	4.1387e-12	1.7321e-06
C_2	3.7928e-25	5.5312e-28	0.0584	4.1667e-33	-	-	0.05838	-
C_3	1.3458	0.3808	8.6087e-51	2	1.3458	0.38081	-	2
C_4	1.3707e-24	1.3369e-30	6.3800e-152	1e-20	-	-	-	-
C_5	4.9532e-40	1.4724e-47	1e-20	4.8225e-75	-	-	-	-
surfaces								
S	0.3907	7.9128	2.9332e-24	1e-20	0.39074	7.9128	-	-
CS_2	0.3046	1.0436	1.2687e-97	6e-29	0.30463	1.0436	-	-
CS_1	9.9968e-21	1e-20	9.9971e-21	1.3889e-59	-	-	-	-

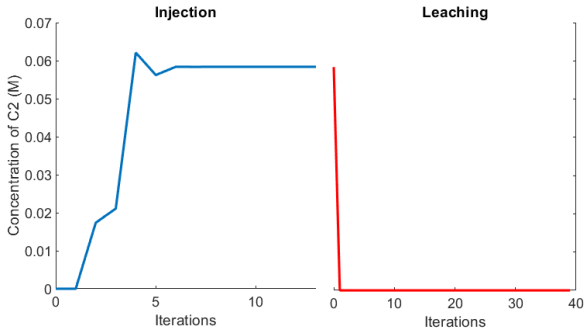

FIGURE 1.23 – Cas test "1D Easy de benchmark MoMas : courbe d'élution pour l'espèce C_2 pendant les périodes d'injection et de lessivage ; équilibre chimique par AA ($m = 2$), MPE et RRE.

L'avantage le plus important des méthodes AA, MPE et RRE par rapport aux méthodes de type Newton-Raphson est que leurs algorithmes ne nécessitent pas le calcul de la matrice jacobienne de la fonction $\mathbf{G}$ (ou $\tilde{\mathbf{G}}$). En résolvant de petits systèmes linéaires par l'algorithme de Newton-Raphson, on rencontre des problèmes numériques dûs au fait que les matrices jacobiennes sont très mal conditionnées [38]. En particulier, pour le test de l'acide gallique (respectivement test "1D Easy" de benchmark MoMas), le conditionnement de la matrice jacobienne varie entre $10^{0.61}$ et $10^{12.6}$ (respectivement entre $10^{3.44}$ et $10^{37.7}$), or comme on a vu déjà avec la méthode d'Accélération d'Anderson, on s'interesse seulement au conditionnement de la matrice $\mathcal{F}_k$ (ou R_k) qu'on obtient toujours inférieur

à 10^{10} grâce à une stratégie de surveillance bien adaptée et des mises à jour des factorisations QR. Pour les deux méthodes polynomiales MPE et RRE, leur efficacité vient du fait qu'elles peuvent être appliquées directement à la solution de systèmes linéaires et non linéaires puisqu'ils ne nécessitent pas une connaissance explicite de la manière dont la séquence est générée.

À ce stade, nous avons atteint le but de la première partie de cette thèse portant sur la résolution numérique des équilibres thermodynamiques en illustrant la grande efficacité de l'approche utilisée (MPE, RRE et AA) par rapport aux résolutions classiques résultant des méthodes de types Newton-Raphson.

Dans la suite de ce travail, nous nous intéressons à l'aspect "écoulement engendrant le transport" de la problématique des transports réactifs. Nous allons présenter l'étude mathématique d'une classe de modèles décrivant des écoulements en eaux peu profondes en interaction avec les eaux de surfaces.

1.4 ÉCOULEMENT DANS DES AQUIFÈRES PEU PROFONDS

Les aquifères sont souvent caractérisés par une forme de stratification des écoulements qui permet de définir des interfaces. La lenteur de la dynamique naturelle assure que ces interfaces sont régulières et ont un comportement stable. Par ailleurs, en raison des dimensions de l'aquifère, l'écoulement est supposé essentiellement orthogonal aux équipotentiels (hypothèse de Dupuit), ce qui permet l'intégration verticale de l'équation de Richards au moins dans la zone saturée. Dans cet esprit, de nombreux modèles 2D ont été développés et utilisés depuis les années 60 (voir par exemple les travaux de Jacob Bear [42, 43]). Une des principales faiblesses de l'approche par intégration verticale réside dans sa justification. Elle n'est valable que pour des échelles de durée et de temps très précises, l'échelle de temps en particulier étant complètement différente des durées typiques des réactions chimiques (voir encore [42] pour les arguments empiriques et qualitatifs, voir [65] pour les calculs asymptotiques). Cependant, de tels modèles $2D$ sont largement utilisés, même hors de leur plage de validité et même s'il s'avère particulièrement difficile de les coupler correctement avec l'écoulement dans la partie non saturée du sous-sol. Une classe de modèles est proposée dans [48] qui consiste à coupler des modèles purement verticaux (pour décrire l'écoulement à petite échelle de temps) avec un modèle horizontal (décrivant l'écoulement à longue échelle de temps). Ils admettent le même comportement asymptotique que le modèle Richards $3D$ à toute échelle de temps lorsque l'aquifère présente une faible profondeur par rapport à ses dimensions horizontales supposées grandes. Ils décrivent l'écoulement essentiellement horizontal dans la partie non saturée de la nappe phréatique ainsi que celui essentiellement vertical dans la partie saturée de l'aquifère.
Dans ce travail, nous étudions deux choix de modèles de type Richards-Dupuit. Pour le premier, l'équation complète de Richards $3D$ est considérée dans la frange capillaire tandis qu'une moyenne verticale de la loi de conservation de masse est faite dans la zone saturée de l'aquifère. Pour le deuxième, nous présentons un nouveau modèle couplé qui est, d'un point de vue physique, très proche du modèle présenté dans [48]. Ce dernier exploite la faible épaisseur d'un aquifère confiné ou non confiné et associe les composantes rapides et lentes de l'écoulement qui sont dominantes dans les aquifères peu profonds. Les deux choix considérés diffèrent d'une part, par la forme de la conductivité hydraulique dans la partie saturée du domaine et d'autre part, selon qu'ils conservent la masse d'eau totale dans le réservoir ou pas.

Géométrie de l'aquifère On considère un aquifère représenté par un domaine multidimensionnel $\Omega := \Omega_x \times (h_{\text{bot}}, h_{\text{soil}})$, où Ω_x ($x = (x_1, x_2)$) est un domaine ouvert dans $\mathbb{R}^n$ ($n \geq 2$) qui correspond à la projection de Ω sur le plan horizontal. On désigne par $\partial\Omega_x$ la frontière de Ω_x. Les bases inférieures et supérieures de Ω sont définies respectivement par les graphes des fonctions $h_{\text{bot}} = h_{\text{bot}}(x)$ et $h_{\text{soil}} =$

$h_{\text{soil}}(x)$ pour $x \in \Omega_x$. On suppose que :

$$h_{\text{soil}}(x) > h_{\text{bot}}(x) \quad \forall x \in \Omega_x. \tag{1.83}$$

Plus précisément, ce domaine est donnée par :

$$\Omega = \{(x,z) \in \Omega_x \times \mathbb{R} \mid z \in]h_{\text{bot}}(x), h_{\text{soil}}(x)[\}. \tag{1.84}$$

On note par $\vec{\nu}$ la normale unitaire extérieure et par $\vec{e}_3$ le vecteur vertical unitaire pointant vers le haut. On décompose la frontière $\partial\Omega$ de Ω en trois zones (bas, haut et latéral) comme suit :

$$\partial\Omega = \Gamma_{\text{bot}} \sqcup \Gamma_{\text{soil}} \sqcup \Gamma_{\text{ver}},$$

avec

$$\Gamma_{\text{bot}} = \{(x,z) \in \Omega \mid z = h_{\text{bot}}(x)\}, \quad \Gamma_{\text{soil}} = \{(x,z) \in \Omega \mid z = h_{\text{soil}}(x)\}, \quad \text{et} \quad \Gamma_{\text{ver}} = \{(x,z) \in \Omega \mid x \in \partial\Omega_x\}.$$

Dans le présent travail, nous dérivons des modèles pour lesquels il est nécessaire de présenter deux sous-régions auxiliaires de Ω (éventuellement dépendant du temps) dans lesquelles l'écoulement présentera un comportement très différent. Soit h la profondeur de l'interface libre séparant la couche d'eau douce et la partie insaturée de l'aquifère, alors la définition de ces sous-régions est basée sur celle de la fonction $h = h(t,x)$ qui pourra être l'une des inconnues de notre modèle. Pour une fonction donnée $h = h(t,x)$ telle que $h_{\text{bot}} \leq h \leq h_{\text{soil}}$, on introduit respectivement les deux sous-régions supérieure et inférieure à $h = h(t,x)$ comme suit :

$$\Omega_t := \{(x,z) \in \Omega \mid z > h(t,x)\} \quad \text{et} \quad \Omega_t^- := \{(x,z) \in \Omega \mid z < h(t,x)\}, \tag{1.85}$$

et

$$\Gamma_t := \{(x,z) \in \Omega \mid z = h(t,x)\}. \tag{1.86}$$

Pour une description tridimensionnelle, on note $\mathbf{x} := (x,z)$ où $x = (x_1, x_2) \in \mathbb{R}^2$ et $z \in \mathbb{R}$ sont les coordonnées usuelles.

Nous allons rappeler tout d'abord l'obtention des équations de Richards ainsi que le principe de l'approximation de Dupuit.

1.4.1 Lois de conservations

Dans un premier temps, on introduit les équations fondamentales couramment utilisées en hydrogéologie ainsi que les paramètres physiques impliqués dans ces équations.
L'équation de Richards représente le mouvement d'un fluide en milieu poreux. Bien qu'attribuée à Lorenzo A. Richards qui l'a publiée en 1931, il est établi que cette équation a en fait été découverte 9 ans plus tôt par Lewis Fry Richardson dans son livre "Weather prediction by numerical process" publié en 1922 [44].

1.4.1.1 Loi de Darcy

Il s'agit d'une loi physique, découverte par l'ingénieur français Henry Darcy en 1856, qui constitue la base fondamentale du calcul de quantité d'eau souterraine ou débit traversant un milieu poreux, par l'hydrodynamique souterraine. Cette loi découle du principe de la conservation du moment cinétique.

Compte tenu des grandes dimensions d'un aquifère par rapport à la taille caractéristique de la structure poreuse du sous-sol, on considère une description continue du milieux poreux.
La vitesse effective q de l'écoulement est donc liée à la pression P par la loi de Darcy donnée par

$$q = \frac{\kappa K_0}{\mu}(\nabla P + \rho g \nabla z), \tag{1.87}$$

où ρ et μ sont respectivement la masse volumique (ou densité) et la viscosité du fluide, z est la hauteur, K_0 est la perméabilité du sol, κ est la conductivité relative et g est la constante d'accélération gravitationnelle.

Notons que l'énergie due à la pression du fluide sur les pores et l'énergie potentielle gravitationnelle sont respectivement représentées par les deux termes ∇P et ∇z. L'énergie cinétique est négligée en raison du mouvement souvent très lent des eaux souterraines.

En introduisant la charge hydraulique H définie par

$$H = \frac{P}{\rho g} + z, \tag{1.88}$$

on peut alors reécrire la loi de Darcy comme suit :

$$q = -K\nabla H - \frac{\kappa K_0}{\mu}(\rho - \rho_0)\, g \nabla z, \tag{1.89}$$

où ρ_0 désigne la densité de référence du fluide.
Dans cette relation, K est la conductivité hydraulique non linéaire qui exprime la capacité du sol à conduire le fluide. Elle est donnée par

$$K = \frac{\kappa \boldsymbol{K}_0 \rho_0\, g}{\mu},$$

où $\boldsymbol{K}_0$ est le tenseur de perméabilité défini ci-dessous.

Tenseur de perméabilité $\boldsymbol{K}_0$: Les propriétés de transmission du sol sont caractérisées par la fonction de porosité ϕ et le tenseur de perméabilité $\boldsymbol{K}_0(x,z)$ qui est un tenseur 3×3 symétrique défini positif décrivant la conductivité du sol *saturé* à la position $(x,z)\in\Omega$. On introduit $\boldsymbol{K}_{xx}\in\mathcal{M}_{22}(\mathbb{R})$, $\boldsymbol{K}_{zz}\in\mathbb{R}^*$ et $\boldsymbol{K}_{xz}\in\mathcal{M}_{21}(\mathbb{R})$ tels que :

$$\boldsymbol{K}_0 = \begin{pmatrix} \boldsymbol{K}_{xx} & \boldsymbol{K}_{xz} \\ \boldsymbol{K}_{xz}^T & \boldsymbol{K}_{zz} \end{pmatrix} \tag{1.90}$$

1.4.1.2 Loi de conservation de la masse du fluide

La loi de conservation de la masse du fluide pendant le déplacement s'écrit comme suit :

$$\partial_t(\theta\rho) + \nabla\cdot(\rho q) = \rho Q, \tag{1.91}$$

où Q désigne un terme source générique (pour le pompage et/ou le réapprovisionnement en eau).
La fonction θ désigne la teneur en humidité volumétrique. Elle est définie par :

$$\theta = \phi\, s, \tag{1.92}$$

où ϕ et s désignent respectivement la porosité du milieu et la saturation. Si on suppose que l'air présent dans la zone non saturée a une mobilité infinie, la saturation s et la fonction θ sont considérées comme des fonctions monotones de la pression comme on le détaillera plus loin.

1.4.1.3 Équation d'état pour la compressibilité du fluide

On considère que le fluide est compressible en supposant que la pression P est liée à la masse volumique par l'équation d'état suivante :

$$\frac{d\rho}{\rho} = \alpha_P \, dP, \tag{1.93}$$

où le réel $\alpha_P \geq 0$ est le coefficient de compressibilité du fluide résultant de la variation de la pression. En intégrant (1.93) on obtient :

$$\rho = \rho_0 \, e^{\alpha_P (P - P_0)}, \tag{1.94}$$

P_0 étant la pression de référence. Si on suppose $\alpha_P = 0$, nous retrouvons le cas incompressible.

1.4.1.4 Hypothèses simplificatrices

Nous donnons dans cette sous-section les hypothèses sur les caractéristiques du fluide et du milieu, mais aussi sur l'écoulement qui sont simplificatrices et significatives dans le contexte de nos problèmes.

Compressibilité du sol On néglige dans le modèle les effets de la compressibilité de la roche, ainsi la porosité ϕ du milieu ne dépend pas des variations de pression et elle est donc supposée être constante.

Compressibilité du fluide Tout d'abord, on suppose que le fluide (à savoir ici l'eau douce) est faiblement compressible. Cela signifie que :

$$\alpha_P \ll 1. \tag{1.95}$$

Exploitons maintenant cette hypothèse : dans des conditions naturelles et en particulier dans un aquifère, on observe une faible mobilité du fluide (définie par le rapport κ/μ). La première conséquence de la faible compressibilité du fluide combinée à cette faible mobilité se traduira dans l'équation du moment cinétique.
Ainsi, en faisant un développement de Taylor par rapport à P de la masse volumique ρ dans le terme de gravité de l'équation de Darcy et en négligeant les termes pondérés par $\alpha_P \, \kappa/\mu \ll 1$ dans (1.89), on obtient :

$$q = -K \nabla H, \quad K = \frac{\kappa(P) \rho_0 \, g}{\mu} \boldsymbol{K}_0. \tag{1.96}$$

La deuxième conséquence, parfois considérée comme une hypothèse supplémentaire appelée *hypothèse de Bear* [45], consiste à négliger les variations de la densité dans la direction de l'écoulement, c'est à dire $\nabla \rho \cdot q \ll 1$. L'équation de la conservation de masse (1.91) s'écrit donc sous la forme simplifiée suivante :

$$\rho \, \partial_t \theta + \theta \partial_t \rho + \rho \nabla \cdot q = \rho Q.$$

Compte tenu de (1.93), c'est à dire $\partial_t \rho = \rho \, \alpha_P \, \partial_t P$, l'équation précédente devient :

$$\rho \, \partial_t \theta + \rho \, \theta \, \alpha_P \, \partial_t P + \rho \nabla \cdot q = \rho Q.$$

On obtient finalement après simplification par $\rho > 0$:

$$\partial_t \theta + \theta \, \alpha_P \, \partial_t P + \nabla \cdot q = Q. \tag{1.97}$$

En tenant compte de la loi de Darcy, l'équation (1.97) peut aussi s'écrire :

$$\partial_t \theta + S_0 \, \partial_t H - \nabla \cdot (K \nabla H) = Q \quad \text{avec} \quad S_0 = \rho_0 \, g \, \phi \, \alpha_P. \tag{1.98}$$

On remarque que si le fluide est supposé incompressible, $\alpha_P = 0$, alors l'équation (1.97) correspond à l'équation de Richards classique en formulation pression. Une définition adéquate de la teneur volumétrique en humidité θ ainsi que de celle de la mobilité κ constituent la clé du modèle.

Hypothèse de Richards Le modèle de Richards est en outre basé sur l'hypothèse que la pression de l'air dans le sous-sol est égale à la pression atmosphérique, ce n'est donc pas une inconnue du problème. On suppose que la saturation et la conductivité relative du sol sont données en fonction de la pression du fluide P, notées respectivement $s = s(P)$ et $\kappa = \kappa(P)$. On introduit la pression de saturation P_s qui est un nombre réel fixe pour définir les trois zones suivantes :

- la zone complètement saturée du support : $\{\mathbf{x} \in \Omega \mid P(.,\mathbf{x}) > P_s\}$;
- la zone partiellement saturée du support : $\{\mathbf{x} \in \Omega \mid P_d < P(.,\mathbf{x}) \leq P_s\}$;
- La zone sèche : $\{\mathbf{x} \in \Omega \mid P(.,\mathbf{x}) \leq P_d\}$.

La teneur en humidité θ est telle que :

$$\theta = \begin{cases} \phi & \text{(zone saturée)} & \text{si} & P(.,\mathbf{x}) > P_s \\ \theta(P) & (0 \leq \theta(P) \leq \phi \text{ et } \theta'(P) > 0) & \text{si} & P_d < P(.,\mathbf{x}) \leq P_s \\ \theta_0 = \phi\, s_0 & \text{(zone sèche)} & \text{si} & P(.,\mathbf{x}) \leq P_d \end{cases} \tag{1.99}$$

où $s_0 > 0$ correspond à une saturation résiduelle positive. La mobilité hydraulique relative associée est alors définie par :

$$\kappa(P) = \begin{cases} 1 & \text{(zone saturée)} & \text{si} & P(.,\mathbf{x}) > P_s \\ \kappa(\theta(P)) & (0 \leq \kappa(P) \leq \phi \text{ et } (\kappa \circ \theta)'(P) > 0) & \text{si} & P_d < P(.,\mathbf{x}) \leq P_s \\ 0 & \text{(zone sèche)} & \text{si} & P(.,\mathbf{x}) \leq P_d \end{cases} \tag{1.100}$$

Il existe un large choix de modèles possibles pour les fonctions s et κ. Les exemples les plus classiques d'un système air-eau sont donnés par le modèle de van Genuchten [46] sans dépendance explicite de la pression de saturation mais avec des paramètres d'ajustement, et le modèle de Brooks et Corey [47]. Le point le plus important qui caractérise ces modèles est que la saturation et la mobilité satisfassent :

$$s(P) = 1 \Longleftrightarrow P \geq P_s \quad \text{et} \quad \kappa(P) = 1 \Longleftrightarrow P \geq P_s. \tag{1.101}$$

En particulier, ces équivalences signifient que la pression de l'eau est supérieure à la pression de saturation P_s si et seulement si le sol est complètement saturé.

1.4.2 Modèle couplant le flux de Richards 3D et le flux horizontal de Dupuit

Dans cette sous-section, on va présenter brièvement un modèle appartenant à la classe de modèles "Dupuit-Richards". En effet, l'équation de Richards $3D$ est considérée dans la frange capillaire tandis qu'une moyenne verticale de la loi de conservation de la masse est faite dans la zone saturée de l'aquifère. Ce modèle diffère légèrement de celui introduit dans [48] parce que l'on considère les équations de Richards $3D$ complètes dans la partie non saturée et non plus uniquement la composante verticale de l'écoulement. Mais la principale différence réside dans la prise en compte de la faible compressibilité du fluide. Cela fera, non seulement, apparaître une transformation bijective qui permettra de neutraliser la nonlinéarité présente dans la dérivée en temps de l'équation de Richards mais aussi de traiter la dégénérescence présente dans l'équation parabolique régissant l'écoulement horizontal de la partie saturée de l'aquifère. En revanche, nous utilisons le même couplage des flux entre les deux zones du réservoir que dans [48]. Il résulte de la propriété de continuité de la composante normale du flux à l'interface de saturation, garantissant ainsi la conservation de la masse des modèles présentés dans [48].

1.4.2.1 Équation de Richards 3D dans la frange capillaire supérieure

Dans la partie non saturée de l'aquifère, Ω_t, l'équation de Richards $3D$ (1.97) donne :

$$\begin{cases} \partial_t\theta + \theta\,\alpha_P\,\partial_t P + \nabla\cdot q = Q & \text{pour } (t,x,z)\in(0,T)\times\Omega_t,\\ q\cdot\vec{\nu} = 0 & \text{pour } (t,x,z)\in(0,T)\times(\Gamma_{\text{soil}}\cup\Gamma_{\text{ver}}),\\ P(t,x,h(t,x)) = P_s & \text{pour } (t,x)\in(0,T)\times\Omega_x,\\ P(0,x,z) = P_{\text{init}}(x,z) & \text{pour } (x,z)\in\Omega_0. \end{cases} \tag{1.102}$$

Cette équation de Richards dépend par définition de la profondeur h qui devrait appartenir à l'intervalle $(h_{\text{bot}}, h_{\text{soil}})$. Nous rappelons que la vitesse effective q est donnée par :

$$q = -K\nabla\left(\frac{P}{\rho_o\, g} + z\right), \quad K = \frac{\kappa(P)\,K_0\rho_0 g}{\mu}.$$

1.4.2.2 Écoulement de Dupuit horizontal dans la zone saturée

Approximation de Dupuit (approche hydrostatique) Cette hypothèse est introduite pour moyenner le problème $3D$ en un problème $2D$ dans la zone saturée du domaine.
L'hypothèse de Dupuit consiste à considérer que la charge hydraulique est constante le long de chaque direction verticale (équipotentiels verticaux). C'est légitime puisque l'on observe effectivement des déplacements quasi horizontaux lorsque l'épaisseur de l'aquifère est faible par rapport à sa largeur et à sa longueur et lorsque l'écoulement est loin des puits et des sources.

Procédure de mise à l'échelle On utilise maintenant les approximations introduites dans 1.4.1.4 pour intégrer verticalement l'équation (1.98), réduisant ainsi le problème tridimensionnel à un problème bidimensionnel. Cette intégration est effectuée entre les profondeurs h_{bot} et h. Comme $\theta(P) = \phi$ dans la zone saturée, la moyenne verticale de (1.98) donne :

$$\int_{h_{\text{bot}}}^{h} (S_0\,\partial_t H + \nabla\cdot q)\,dz = \int_{h_{\text{bot}}}^{h} Q\,dz. \tag{1.103}$$

On désigne par $B_f = h - h_{\text{bot}}$ l'épaisseur de la zone saturée et par $\tilde{Q}$ le terme source représentant l'approvisionnement moyenné d'eau douce dans l'aquifère, on a :

$$\tilde{Q} = \frac{1}{B_f}\int_{h_{\text{bot}}}^{h} Q\,dz.$$

En appliquant la formule de Leibniz au premier au membre gauche de l'égalité (1.103), on obtient :

$$\int_{h_{\text{bot}}}^{h} S_0\,\partial_t H\,dz = S_0\,\frac{\partial}{\partial t}\int_{h_{\text{bot}}}^{h} H\,dz - S_0\,H_{|z=h}\partial_t h + S_0\,H_{|z=h_{\text{bot}}}\,\partial_t h_{\text{bot}}.$$

On introduit la moyenne verticale de la charge hydraulique, notée $\tilde{H}$, comme suit :

$$\tilde{H} = \frac{1}{B_f}\int_{h_{\text{bot}}}^{h} H\,dz.$$

D'après l'approximation de Dupuit qui implique que $H(x_1,x_2,z) \simeq \tilde{H}(x_1,x_2)$ pour $x=(x_1,x_2)\in\Omega$ et $z\in(h_{\text{bot}},h)$, on déduit que :

$$\int_{h_{\text{bot}}}^{h} S_0\,\partial_t H\,dz = S_0\,B_f\,\partial_t\tilde{H}.$$

De même, on a :

$$\int_{h_{\text{bot}}}^{h} \nabla \cdot q \, dz = \nabla' \cdot (B_f \, \tilde{q'}) + q_{|z=h^-} \cdot \nabla(z-h) - q_{|z=h_{\text{bot}}^+} \cdot \nabla(z-h_{\text{bot}}),$$

où $\nabla' = (\partial_{x_1}, \partial_{x_2})$, $q' = (q_{x_1}, q_{x_2})$ et $\tilde{q'} = \frac{1}{B_f} \int_{h_{\text{bot}}}^{h} q' \, dz$ est la vitesse moyenne de Darcy.
Puisque $\kappa(P) = 1$ pour $z \in (h_{\text{bot}}, h)$, $\tilde{q'}$ est donc donnée par :

$$\tilde{q'} = -\frac{1}{B_f} \int_{h_{\text{bot}}}^{h} (K \nabla' H) \, dz = -\frac{1}{B_f} \int_{h_{\text{bot}}}^{h} (K \nabla' \tilde{H}) \, dz = -\tilde{K} \nabla' \tilde{H}, \quad \tilde{K} = \frac{1}{B_f} \int_{h_{\text{bot}}}^{h} \frac{K_0 \, \rho_0 \, g}{\mu} \, dz.$$

Finalement, la loi de conservation de la masse d'eau douce, moyennée dans la zone saturée peut s'écrire comme ci-dessous : suit :

$$S_0 \, B_f \, \partial_t \tilde{H} = \nabla' \cdot (B_f \, \tilde{K} \nabla' \tilde{H}) + q_{|z=h_{\text{bot}}^+} \cdot \nabla(z-h_{\text{bot}}) - q_{|z=h^-} \cdot \nabla(z-h) + B_f \, \tilde{Q}. \tag{1.104}$$

Dans cette équation, le terme $B_f \, \tilde{K}$ peut être considéré comme étant la transmissivité dynamique de la couche d'eau douce. À ce stade, on a obtenu un système de deux équations aux dérivées partielles ((1.97),(1.104)) à trois inconnues $P, \tilde{H}$ et h.

Flux et équations de continuité à travers l'interface Notre objectif est maintenant d'inclure dans le modèle les propriétés de continuité et de transfert à travers l'interface afin d'exprimer les deux termes de flux apparaissant dans (1.104) et réduire ainsi le nombre d'inconnues.

1. **Flux à travers l'interface de saturation :** L'interface de saturation est caractérisée par l'équation cartésienne $F(x_1, x_2, z, t) = 0 \Leftrightarrow z - h(x_1, x_2, t) = 0$, le vecteur normal unitaire à l'interface $\vec{\nu}$ est donc colinéaire à $\nabla(z-h)$.
   La relation régissant la continuité de la composante normale de la vitesse s'écrit alors :

   $$(q_{|z=h^+} - q_{|z=h^-}) \cdot \vec{\nu} = 0 \Leftrightarrow q_{|z=h^+} \cdot \nabla(z-h) = q_{|z=h^-} \cdot \nabla(z-h). \tag{1.105}$$

2. **Approximation du flux** $q_{|z=h^+} \cdot \nabla(z-h)$ **:** Le flux $q_{|z=h^+} \cdot \nabla(z-h)$ exprime les transferts de masse entre les deux parties de l'aquifère. Comme dans [48], on approche ce flux par :

   $$q_{|z=h^+} \cdot \nabla(z-h) \simeq \int_{h(t,x)}^{h_{\text{soil}}(x)} (\phi \frac{\partial s(P)}{\partial t} + \phi \, s(P) \, \alpha_P \frac{\partial P}{\partial t} - Q) \, dz. \tag{1.106}$$

   Cette approximation provient de l'hypothèse d'une conductivité hydraulique horizontale quasi nulle (i.e. $K_{xx} \simeq (0)$) dans la frange capillaire. Ceci correspond à un écoulement presque vertical dans cette partie de l'aquifère. L'équation de Richards $3D$ est donc réduite à une équation $1D$ dont l'intégration entre h et h_{soil} donne l'approximation (1.106).
   Ceci constitue une différence essentielle avec l'analyse mathématique présentée dans [49] dans laquelle les échanges entre les deux parties de l'aquifère ont été simplifiés et représentés par l'ajout d'un terme source externe, découplant ainsi les deux problèmes.

3. **Couche imperméable en** $z = h_{\text{soil}}$: Puisque la couche inférieure est imperméable, il n'y a pas de flux à travers la frontière $z = h_{\text{bot}}$,d'où :

   $$q(h_{\text{bot}}) \cdot \nabla(z - h_{\text{bot}}) = 0. \tag{1.107}$$

4. **Équations de continuité :** La relation de continuité imposée à l'interface permet de réduire correctement le nombre d'inconnues dans les équations ((1.97),(1.104)).
   L'approximation de Dupuit se traduit par $\tilde{H} = H_{|z=h^-}$, la pression P satisfait alors dans Ω_t^- :

   $$P(t,x,z) = \rho_0 \, g \, (\tilde{H}(t,x) - z), \quad t \in [0, T[\quad \text{et} \quad (x,z) \in \Omega_t^-. \tag{1.108}$$

   En outre, la pression est continue à travers l'interface Γ_t, il s'en suit que :

   $$P(t,x,h^-) = P(t,x,h^+) = P_s \Leftrightarrow \tilde{H} = \frac{P_s}{\rho_0 \, g} + h. \tag{1.109}$$

D'après l'équation (1.109), on peut remplacer $\tilde{H}$ par h dans (1.104), on obtient alors :

$$S_0\, B_f\, \partial_t h - \nabla' \cdot (B_f\, \tilde{K} \nabla' h) = B_f\, \tilde{Q} - \int_{h(t,x)}^{h_{\text{soil}}(x)} (\phi \frac{\partial s(P)}{\partial t} + \phi\, s(P)\, \alpha_P \frac{\partial P}{\partial t} - Q)\, dz \quad \text{dans } (0,T) \times \Omega_x, \tag{1.110}$$

$$\tilde{K} \nabla' h \cdot \vec{\nu} = 0 \quad \text{sur } (0,T) \times \partial\Omega_x.$$

avec

$$B_f = (h - h_{\text{bot}}), \quad \tilde{K} = \frac{1}{B_f} \int_{h_{\text{bot}}}^{h} \frac{K_0\, \rho_0\, g}{\mu}\, dz \quad \text{et} \quad S_0 = \rho_0\, g\, \phi\, \alpha_P. \tag{1.111}$$

On impose la condition de Neumann homogène sur $\partial\Omega_x$ afin de simplifier la présentation.

Conclusion Le modèle final $(\mathcal{M})$ couplant l'équation de Richards $3D$ (pour la description de l'écoulement dans la frange capillaire) et le flux horizontal de Dupuit consiste en les systèmes (1.102),(1.108) et (1.110) :

- Dans Ω_t, l'équation de Richards $3D$ (1.97) donne :

$$\begin{cases} \partial_t \theta + \theta\, \alpha_P\, \partial_t P + \nabla \cdot q = Q & \text{dans } (0,T) \times \Omega_t, \\ q \cdot \vec{\nu} = 0 & \text{sur } (0,T) \times (\Gamma_{\text{soil}} \cup \Gamma_{\text{ver}}), \\ P(t,x,h(t,x)) = P_s & \text{dans } (0,T) \times \Omega_x, \\ P(0,x,z) = P_{\text{init}}(x,z) & \text{dans } \Omega_0. \end{cases} \tag{1.112}$$

  La vitesse effective q est donnée par :

$$q = -K \nabla (\frac{P}{\rho_0\, g} + z) \quad \text{avec} \quad K = \frac{\kappa(P)\, K_0\, \rho_0\, g}{\mu}.$$

- Dans Ω_t^-, la pression P satisfait :

$$P(t,x,z) = \rho_0\, g\, (\frac{P_s}{\rho_0\, g} + h - z) \quad \text{dans} \quad (0,T) \times \Omega_t^-.$$

- La profondeur de l'interface Γ_t, h, satisfait dans Ω_x :

$$\begin{cases} S_0\, B_f\, \partial_t h - \nabla' \cdot (B_f\, \tilde{K} \nabla' h) = B_f\, \tilde{Q} - \int_{h(t,x)}^{h_{\text{soil}}(x)} (\phi \frac{\partial s(P)}{\partial t} + \phi\, s(P)\, \alpha_P \frac{\partial P}{\partial t} - Q)\, dz & \text{dans } (0,T) \times \Omega_x, \\ \tilde{K} \nabla' h \cdot \vec{\nu} = 0 & \text{sur } (0,T) \times \partial\Omega_x, \\ h(0,x) = h_0(x) & \text{dans } \Omega_x. \end{cases} \tag{1.113}$$

1.4.3 Modèle couplant la composante rapide et lente de l'écoulement dans des aquifères peu profonds.

Dans cette section nous rappelons une classe de modèles présentée dans [48]. L'objectif principal étant de donner un modèle proche du modèle originel, Richards $3D$, tout en étant plus simple à traiter numériquement. Dans chaque modèle de cette classe, nous considérons un aquifère occupant un domaine géométrique peu profond par rapport à ses dimensions horizontales. Il s'avère que deux types d'écoulement dominant se superposent dans ce type d'aquifères peu profonds. Ils sont classiquement modélisés par des systèmes mathématiques aux structures très différentes. Le premier correspond à la composante rapide de l'écoulement et a lieu principalement dans la direction verticale et dans la partie insaturée de l'aquifère. Le second correspond à la composante plus lente de

l'écoulement et a lieu globalement dans la direction horizontale et dans la partie saturée de l'aquifère. En particulier, l'écoulement vertical apparaît comme étant instantané dans cette zone.
La classe de modèles que nous présentons ici, est basée sur le couplage de ces deux types d'écoulement.

Le tenseur de perméabilité $\boldsymbol{K}_0$ donné par (1.90).

Conditions aux limites Pour prendre en compte l'écoulement venant ou allant vers la surface du sol, on considère une condition générale de Robin sur la frontière Γ_{soil} :

$$a\,P + q\cdot\vec{\nu} = F \quad \text{pour } (t,x,z)\in(0,T)\times\Gamma_{\text{soil}}, \tag{1.114}$$

avec $a\in\mathbb{R}$ et F un terme source. En revanche, on suppose que la couche du fond de l'aquifère Γ_{bot} est imperméable. Pour simplifier la présentation, on prend également une telle condition sur la partie latérale Γ_{ver} :

$$q\cdot\vec{\nu} = 0 \quad \text{pour } (t,x,z)\in(0,T)\times\Gamma_{\text{bot}}\cup\Gamma_{\text{ver}}. \tag{1.115}$$

Avant de présenter le modèle que nous allons étudier dans cette partie, nous rappelons les équations de Richards $3D$ classiquement utilisées pour décrire l'écoulement de l'eau dans un aquifère ainsi que la variante compressible des équations de Richards $3D$ que nous avons détaillées précédemment (paragraphe 1.4.1).

Problème de Richards $3D$

$$\begin{cases} \partial_t\theta(P)+\nabla\cdot\boldsymbol{v}=0 & \text{dans } (0,T)\times\Omega \\ \boldsymbol{v}=-\kappa(P)\,\boldsymbol{K}_0\left(\dfrac{1}{\rho g}\nabla P+\boldsymbol{e}_3\right) & \text{dans } (0,T)\times\Omega \\ \alpha\,P+\beta\,\boldsymbol{v}\cdot\boldsymbol{n}=F & \text{sur } (0,T)\times\Gamma_{\text{soil}} \\ \boldsymbol{v}\cdot\boldsymbol{n}=0 & \text{sur } (0,T)\times(\Gamma_{\text{bot}}\cup\Gamma_{\text{ver}}) \\ P(0,x,z)=P_{\text{init}}(x,z) & \text{pour } (x,z)\in\Omega \end{cases} \tag{1.116}$$

Problème de Richards $3D$ **pour un fluide faiblement compressible**

$$\begin{cases} \partial_t\theta(P)+\theta\alpha_P\,\partial_t P+\nabla\cdot\boldsymbol{v}=0 & \text{dans } (0,T)\times\Omega, \\ \boldsymbol{v}=-\kappa(P)\,\boldsymbol{K}_0\nabla\left(\dfrac{P}{\rho g}+z\right) & \text{dans } (0,T)\times\Omega, \\ P(0,x,z)=P_{\text{init}}(x,z) & \text{for } (x,z)\in\Omega. \\ +\text{conditions aux limites (1.114) et (1.115)}, \end{cases} \tag{1.117}$$

le paramètre α_p caractérisant la compressibilité de l'eau. Une justification de ce modèle est donnée dans [67].

Conductivité hydraulique moyenne On introduit :

$$\boldsymbol{S}_0=\boldsymbol{K}_{xx}-\frac{1}{\boldsymbol{K}_{zz}}\boldsymbol{K}_{xz}(\boldsymbol{K}_{xz}^T) \quad\text{et}\quad \boldsymbol{M}_0=\begin{pmatrix}\boldsymbol{S}_0 & 0\\ 0 & 0\end{pmatrix}. \tag{1.118}$$

Le tenseur $\boldsymbol{S}_0$ de taille 2×2 est le complèment de Schur du bloc $\boldsymbol{K}_{zz}$ dans le tenseur $\boldsymbol{K}_0$. $\boldsymbol{S}_0$ jouera le rôle du tenseur de perméabilité effective. On introduit également le tenseur de conductivité moyen $\tilde{\boldsymbol{K}}$ défini dans $(0,T)\times\Omega_x$ pour toute fonction $\tilde{H}=\tilde{H}(t,x)$ par :

$$\tilde{\boldsymbol{K}}(\tilde{H})(t,x)=\int_{h_{\text{bot}}(x)}^{h_{\text{soil}}(x)}\kappa\Big(\rho g(\tilde{H}(t,x)-z)\Big)\,\boldsymbol{S}_0(x,z)\,dz. \tag{1.119}$$

La classe des modèles

- Dans Ω_t, l'équation $1D$ de Richards suivante est valable :

$$\begin{cases} \partial_t \theta(P) + \partial_z\Big(\boldsymbol{q}\cdot e_3\Big) = 0 & \text{dans } (0,T)\times\Omega_t, \\ a\,P + \boldsymbol{q}\cdot\boldsymbol{e}_3 = F & \text{dans } (0,T)\times\Gamma_{\text{soil}}, \\ P\Big(t,x,h(t,x)\Big) = \rho\, g(\tilde{H}-h) & \text{dans } (0,T)\times\Omega_x, \\ P(0,x,z) = P_{\text{init}}(x,z) & \text{dans } \Omega_0. \end{cases} \tag{1.120}$$

- Dans Ω_t^-, la pression d'eau P satisfait :

$$P(t,x,z) = \rho\, g\Big(\tilde{H}(t,x) - z\Big) \quad \text{pour } t\in[0,T[\,,\ (x,z)\in\Omega_t^-. \tag{1.121}$$

- La charge hydraulique $\tilde{H}$ est une solution dans Ω_x du problème suivant :

$$\begin{cases} -\nabla'\cdot(\tilde{\boldsymbol{K}}\nabla'\tilde{H}) = -(\boldsymbol{q}\cdot\boldsymbol{e}_3)|_h^+ & \text{pour } (t,x)\in]0,T[\times\Omega_x, \\ \tilde{\boldsymbol{K}}(\tilde{H})\,\nabla'\tilde{H}\cdot\vec{\nu} = 0 & \text{pour } (t,x)\in]0,T[\times\partial\Omega_x \\ \tilde{H}(0,x) = H_{\text{init}}(x) & \text{pour } x\in\Omega_x \end{cases} \tag{1.122}$$

  où $(\boldsymbol{q}\cdot e_3)\big|_{\Gamma_h^+}$ désigne la trace de $(\boldsymbol{q}\cdot e_3)$ sur Γ_t par valeurs supérieures.

- Le niveau $z=h$ en dessous duquel l'écoulement vertical est supposé instantanée est défini par :

$$h(t,x) = \max\Big\{\min\Big\{\tilde{H}(t,x) - \frac{P_s}{\rho\, g}, h_{\text{soil}}(x)\Big\}, h_{\text{bot}}(x)\Big\}. \tag{1.123}$$

- La vitesse de l'eau $\boldsymbol{v}$ est définie dans Ω par :

$$\boldsymbol{v} = \boldsymbol{q} + \boldsymbol{w} \qquad \text{pour } t\in]0,T[\,,(x,z)\in\Omega, \tag{1.124}$$

  et pour $t\in]0,T[$, $(x,z)\in\Omega$, les vitesses auxiliaires sont données par :

$$\boldsymbol{q} = -\kappa(P)\,\boldsymbol{K}_{zz}\left(\frac{1}{\rho\, g}\partial_z P + 1\right)\boldsymbol{e}_3, \quad \boldsymbol{w} = -\kappa\Big(\rho\, g(\tilde{H}-z)\Big)\,\boldsymbol{M}_0\nabla'\tilde{H}. \tag{1.125}$$

Nous continuons en donnant un bref aperçu des propriétés de la solution du problème (1.120)–(1.125). Nous renvoyons à [48] pour plus de détails. Ce modèle est une alternative au problème de Richards $3D$ pour décrire l'écoulement dans un aquifère peu profond dans une large gamme d'échelles de temps. La première remarque concernant le modèle (1.120)–(1.125) est qu'il a deux avantages par rapport au modèles de Richards $3D$ dont il est issu :

— Premièrement il engendre des problèmes numériques plus rapides à résoudre. En effet le problème (1.120)–(1.125) est le couplage d'un problème $2D$ avec une multitude de problèmes de Richards $1D$ verticaux qui pourront être résolus en parallèle offrant ainsi un gain de temps important dans le calcul numérique.

— Deuxièment, ce problème couplé et le problème original de Richards $3D$ présentent les mêmes comportements dominants lorsque le *rapport* ϵ = profondeur/largeur de l'aquifère est petit. En effet, il est montré dans [48] que le $3D$-Richards et ce modèle couplé admettent exactement les mêmes comportements asymptotiques lorsque $\epsilon\to 0$. De plus ces comportements effectifs sont identiques pour tous les choix d'échelles de temps considérés. Ces problèmes effectifs sont donnés dans la sous-section 1.4.5 par (1.145)–(1.150).

Le modèle (1.120)–(1.125) est un modèle exprimé dans les variables physiques qui couplent plusieurs types d'écoulements apparaissant dans les problèmes effectifs pour des échelles de temps, en temps court et en temps long (voir (1.145)–(1.150).). Le premier système est un problème $1D$ de Richards vertical dans la partie supérieure de l'aquifère et est associé à la vitesse $\boldsymbol{q}$. Il reproduit le comportement de la solution en temps court. Le second système est un problème $2D$ horizontal qui suppose un écoulement vertical instantané dans la partie inférieure de l'aquifère. La vitesse associée est $\boldsymbol{w}$ et il reproduit le comportement de la solution en temps long.
Le nouveau modèle couplé que nous considérons dans la section 1.4.4 a des propriétés physiques très proches du modèle (1.120)–(1.125), notamment il est conservatif et a les mêmes comportements asymptotiques à différentes échelles de temps. Il sera justifié formellement à l'aide d'arguments d'analyse asymptotique dans la section 1.4.5.

1.4.4 Modèle conservatif anisotrope pour un fluide compressible.

Nous cherchons à prouver l'existence de la solution d'un modèle de type (1.119)-(1.125), introduit dans [48]. Comme indiqué à la fin du le paragraphe suivant, ce problème est difficile à étudier tel quel. Notre but maintenant est donc de proposer un nouveau modèle, physiquement très proche de (1.119)-(1.125), mais plus adapté à l'étude théorique.
Nous présentons et justifions une généralisation du problème (1.120)–(1.125) dans le cadre d'un fluide faiblement compressible (1.117) (voir Sous-section 1.4.1) et dans le cas où la conductivité horizontale n'est plus nulle dans la partie supérieure de l'aquifère Ω_t. Cette composante horizontale du flux dans Ω_t sera caractérisée par le tenseur symétrique défini positif $\boldsymbol{N}_0$ de taille 2×2. On introduit également :

$$\boldsymbol{A}_0=\boldsymbol{M}_0-\begin{pmatrix}\boldsymbol{N}_0 & 0\\ 0 & 0\end{pmatrix}=\begin{pmatrix}\boldsymbol{S}_0-\boldsymbol{N}_0 & 0\\ 0 & 0\end{pmatrix},\quad \boldsymbol{B}_0=\begin{pmatrix}\boldsymbol{N}_0 & 0\\ 0 & \boldsymbol{K}_{zz}\end{pmatrix}\quad\text{et}\quad \boldsymbol{G}_0=\begin{pmatrix}\boldsymbol{N}_0 & 0\\ 0 & 0\end{pmatrix}. \tag{1.126}$$

On considère le paramètre de compressibilité du fluide $\alpha_p>0$ tel que $\alpha_p\ll 1$. Rappelons que, dans ce cadre, les équations de Richards $3D$ compressibles (1.117) ont été obtenues.

Modèle généralisé Le problème généralisé est donné par les équations (1.127)–(1.133) suivantes :

- La vitesse de l'eau $\boldsymbol{v}$ est définie dans Ω par :

$$\boldsymbol{v}=\boldsymbol{q}+\boldsymbol{w}\quad\text{pour } t\in]0,T[\,,(\boldsymbol{x},z)\in\Omega, \tag{1.127}$$

  et pour $t\in]0,T[$, $(\boldsymbol{x},z)\in\Omega$, les vitesses auxiliaires sont données par :

$$\boldsymbol{q}=-\kappa(P)\,\boldsymbol{B}_0\left(\frac{1}{\rho g}\nabla P+\boldsymbol{e}_3\right),\quad \boldsymbol{w}=-\kappa\Big(\rho g(\tilde{H}-z)\Big)\,\boldsymbol{A}_0\nabla\tilde{H}. \tag{1.128}$$

- Dans $\Omega_t(t)$, l'équation Richards $3D$ suivante est valable :

$$\begin{cases}\partial_t\theta(P)+\alpha_P\theta(P)\partial_t P+\nabla\cdot\boldsymbol{q}=0 & \text{pour } t\in]0,T[\,,\ (\boldsymbol{x},z)\in\Omega_t(t)\\ \alpha P+\beta\,\boldsymbol{q}\cdot\boldsymbol{n}=F & \text{pour } (t,\boldsymbol{x},z)\in]0,T[\times\Gamma_{\text{soil}}\\ \boldsymbol{q}\cdot\boldsymbol{n}=0 & \text{pour } (t,\boldsymbol{x})\in]0,T[\times\Gamma_{\text{ver}}\\ P\Big(t,\boldsymbol{x},h(t,\boldsymbol{x})\Big)=\rho g\Big(\tilde{H}(t,\boldsymbol{x})-h(t,\boldsymbol{x})\Big) & \text{pour } (t,\boldsymbol{x})\in]0,T[\times\Omega_x\\ P(0,\boldsymbol{x},z)=P_{\text{init}}(\boldsymbol{x},z) & \text{pour } (\boldsymbol{x},z)\in\Omega_t(0)\end{cases} \tag{1.129}$$

- Dans $\Omega_t^-(t)$, la pression de l'eau P satisfait :

$$P(t,\boldsymbol{x},z)=\rho g\Big(\tilde{H}(t,\boldsymbol{x})-z\Big)\qquad\text{pour } t\in[0,T[\,,\ (\boldsymbol{x},z)\in\Omega_t^-(t) \tag{1.130}$$

- La charge hydraulique $\tilde{H}$ est une solution dans Ω_x du problème suivant :

$$\begin{cases} \rho g \alpha_P (h - h_{\text{bot}}) \partial_t \tilde{H} - \nabla' \cdot \left(\tilde{\boldsymbol{J}}(\tilde{H}) \nabla' \tilde{H} \right) = -(\boldsymbol{q} \cdot (\boldsymbol{e}_3 - \nabla' h))\big|_{\Gamma_h^+} & (t, \boldsymbol{x}) \in]0, T[\times \Omega_x \\ \tilde{\boldsymbol{J}}(\tilde{H}) \nabla' \tilde{H} \cdot \boldsymbol{n} = 0 & (t, \boldsymbol{x}) \in]0, T[\times \partial\Omega_x \\ \tilde{H}(0, \boldsymbol{x}) = H_{\text{init}}(\boldsymbol{x}) & \boldsymbol{x} \in \Omega_x \end{cases} \tag{1.131}$$

  où $\left(\boldsymbol{q} \cdot (\boldsymbol{e}_3 - \nabla' h)\right)\big|_{\Gamma_h^+}$ désigne la trace normale de $\boldsymbol{q}$ sur Γ_t par valeurs supérieures.

- La conductivité hydraulique moyenne est donnée par :

$$\tilde{\boldsymbol{J}}(\tilde{H})(t, \boldsymbol{x}) = \tilde{\boldsymbol{K}}(\tilde{H}) - \int_{h(t,\boldsymbol{x})}^{h_{\text{soil}}(\boldsymbol{x})} \kappa\Big(\rho\, g(\tilde{H}(t, \boldsymbol{x}) - z)\Big)\, \boldsymbol{N}_0(\boldsymbol{x}, z)\, dz. \tag{1.132}$$

- Le niveau $z = h$ en dessous duquel on considère l'écoulement vertical comme instantané est défini tel que :

$$h(t, \boldsymbol{x}) = \max\left\{ \min\left\{ \tilde{H}(t, \boldsymbol{x}) - \frac{P_s}{\rho g}, h_{soil}(\boldsymbol{x}) \right\}, h_{\text{bot}}(\boldsymbol{x}) \right\}. \tag{1.133}$$

Nous supposons que $\boldsymbol{N}_0$ est assez petit pour que $\tilde{\boldsymbol{J}}$ and $\boldsymbol{B}_0$ soient définis positifs.

Commentaires sur le modèle couplé généralisé Nous donnons maintenant quelques commentaires importants sur le problème (1.127)–(1.133), en particulier pour mettre en évidence ses différences avec le problème d'origine (1.120)–(1.125).

— La première différence concerne la compressibilité. Celle-ci apparaît dans la première équation de (1.129). Cela signifie que nous considérons un problème de Richards $3D$ compressible dans la partie supérieure de l'aquifère. La compressibilité a également un impact sur le comportement de la solution dans la partie inférieure de l'aquifère à travers le premier terme de la première équation de (1.131). En particulier le problème (1.127)–(1.133), comme l'était (1.120)–(1.125), est conservatif contrairement au modèle ($\mathcal{M}$).

— Le vecteur vitesse $\boldsymbol{v}$ s'avère être toujours la superposition des deux vitesses $\boldsymbol{q}$ et $\boldsymbol{w}$ (voir les équations (1.124) and (1.127)).

— Contrairement à (1.120)–(1.125), le composant $\boldsymbol{q}$ n'imite pas exactement le comportement de la solution du modèle Richards $3D$ pour une échelle de temps court (problème $1D$-Richards vertical). En effet, la vitesse $\boldsymbol{q}$ contient une composante horizontale non nulle provenant du tenseur $\boldsymbol{G}_0$. Cela signifie que nous considérons maintenant un petit écoulement horizontal dans la partie supérieure de l'aquifère. L'idée est que les composantes horizontales de l'écoulement ne sont pas dominantes dans les aquifères peu profonds pour une échelle de temps court. Cet ajout de $\boldsymbol{G}_0$ dans l'équation est donc possible et n'aura aucun impact sur le problème effectif pour une échelle en temps court. En revanche, les composantes horizontales ne sont pas négligeables pour une échelle en temps long et l'introduction de $\boldsymbol{G}_0$ aura donc un impact sur le problème effectif pour une échelle en temps long.

— Comme dans (1.120)–(1.125), la composante $\boldsymbol{w}$ est horizontale. Néanmoins, elle est caractérisée par le tenseur $\boldsymbol{A}_0$ au lieu de $\boldsymbol{M}_0$. En raison de l'introduction de $\boldsymbol{G}_0$, cette composante de la vitesse n'est plus la seule horizontale qui apparaît dans le problème. Ceci justifie la définition (1.126) de $\boldsymbol{A}_0$.

— Le même genre de différence s'observe pour la conductivité moyenne $\tilde{\boldsymbol{J}}$. En effet, elle prend en compte la composante horizontale de $\boldsymbol{q}$ qui apparaît de Ω_t.

L'objectif de la section suivante est de donner une justification formelle de ce modèle en adimensionnant le problème et en étudiant le comportement asymptotique de la solution pour différentes échelles de temps.

1.4.5 Justification du modèle généralisé et développements asymptotiques formels

Nous allons donc présenter maintenant une justification formelle du modèle (1.127)–(1.133). Plus précisément et suivant la stratégie donnée dans [48], nous déterminons les problèmes effectifs associés à la fois au problème original de Richards $3D$ (1.116) et au problème couplé (1.127)–(1.133). Ces problèmes effectifs déterminent les comportements dominants de l'écoulement dans un aquifère peu profond, et ce, en fonction de l'échelle de temps considérée. Pour l'analyse asymptotique, la question est liée au comportement des modèles sans dimension (1.134) et (1.135)-(1.140). Nous utilisons des arguments d'analyse asymptotique lorsque le rapport *profondeur caractéristique / longueur caractéristique de l'aquifère peu profond* est très faible (tend vers zéro).

On considère un domaine de référence fixe $\overline{\Omega}$ de type (1.84) et un nombre réel $\bar{T} > 0$ sans dimension. On fixe $\overline{\Omega}_x$, $\overline{h}_{\rm soil}$, et $\overline{h}_{\rm bot}$ tels que :

$$\overline{\Omega} = \Big\{ (\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega}_x \times \mathbb{R} \quad | \quad \overline{z} \in \Big] \overline{h}_{\rm bot}(\overline{\boldsymbol{x}}), \overline{h}_{\rm soil}(\overline{\boldsymbol{x}}) \Big[\Big\}.$$

La frontière de $\overline{\Omega}$ est décomposée en : $\overline{\Gamma}_{\rm bot} := \{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega} \mid \overline{z} = \overline{h}_{\rm bot}(\overline{\boldsymbol{x}})\}$, $\overline{\Gamma}_{\rm soil} := \{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega} \mid \overline{z} = \overline{h}_{\rm soil}(\overline{\boldsymbol{x}})\}$, et $\overline{\Gamma}_{\rm ver} := \{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega} \mid \overline{\boldsymbol{x}} \in \partial\overline{\Omega}_x\}$.
On introduit les nombres positifs T, L_x et L_z qui représentent respectivement les temps, largeur et profondeur caractéristiques. Puis on introduit les variables physiques en fonction des variables sans dimension $\overline{\boldsymbol{x}}$, $\overline{z}$ et $\bar{t}$ par :

$$\boldsymbol{x} = L_x \overline{\boldsymbol{x}}, \quad z = L_z \overline{z}, \quad t = T\bar{t}/\bar{T}.$$

On renvoie au chapitre 4 (partie Annexe) pour la définition des différents paramètres physiques sans dimension.

1.4.5.1 Problème adimensionné

Puisque l'aquifère est supposé mince par rapport à sa largeur horizontale, la quantité L_z/L_x est donc faible. On choisit de considérer un aquifère de hauteur fixe d'ordre $L_z = 1$ et de grande dimension horizontale $L_x = 1/\varepsilon$ pour $0 < \varepsilon \ll 1$. Nous donnons maintenant une version "ré-échelonnée" du problème de Richards $3D$ (1.117) et du modèle couplé (1.127)–(1.133).

Problème de Richards $3D$ compressible adimensionné L'équation de conservation de masse associée à la loi de Darcy et aux conditions aux limites permet d'obtenir le problème de Richards $3D$ compressible adimensionné suivant :

$$\begin{cases} \dfrac{\overline{T}}{\bar{T}} \partial_{\bar{t}} \theta(\overline{P}) + \alpha_p \dfrac{\overline{T}}{\bar{T}} s(\overline{P}) \partial_{\bar{t}} \overline{P} + \varepsilon \nabla_{\overline{\boldsymbol{x}}} \cdot (\overline{\boldsymbol{v}}) + \partial_{\overline{z}} \overline{\boldsymbol{v}} \cdot \boldsymbol{e}_3 = 0 & \text{dans }]0, \bar{T}[\times \overline{\Omega} \\ \overline{\boldsymbol{v}} = -\kappa(\overline{P}) \, \overline{\boldsymbol{K}}_0 \left(\dfrac{\varepsilon}{\rho g} \nabla_{\overline{x}} \overline{P} + \Big(\dfrac{1}{\rho g} \partial_{\overline{z}} \overline{P} + 1 \Big) \boldsymbol{e}_3 \right) & \text{dans }]0, \bar{T}[\times \overline{\Omega}, \\ \alpha \, \overline{P} \left(\varepsilon^2 \left\| \nabla_{\overline{x}} \overline{h}_{\rm soil} \right\|^2 + 1 \right)^{1/2} + \beta \, \overline{\boldsymbol{v}} \cdot \left(\boldsymbol{e}_3 - \varepsilon \nabla_{\overline{x}} \overline{h}_{\rm soil} \right) = \left(\varepsilon^2 \left\| \nabla_{\overline{x}} \overline{h}_{\rm soil} \right\|^2 + 1 \right)^{1/2} \overline{F} & \text{sur }]0, \bar{T}[\times \overline{\Gamma}_{\rm soil}, \\ \overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{n}} = 0 & \text{sur }]0, \bar{T}[\times \overline{\Gamma}_{\rm ver}, \\ \overline{\boldsymbol{v}} \cdot \left(\varepsilon \nabla_{\overline{x}} \overline{h}_{\rm bot} - \boldsymbol{e}_3 \right) = 0 & \text{sur }]0, \bar{T}[\times \overline{\Gamma}_{\rm bot}. \end{cases} \tag{1.134}$$

Modèle couplé (1.127)–(1.133) adimensionné Pour le même paramètre $\varepsilon \ll 1$, la version adimensionnée du problème couplé (1.127)–(1.133) est donnée par l'équation de Richards $1D$ dans la zone de

transition

$$\begin{cases} \dfrac{\overline{T}}{T}\partial_{\bar t}\theta(\overline{P})+\alpha_p\dfrac{\overline{T}}{T}s(\overline{P})\partial_{\bar t}\overline{P}+\partial_{\bar z}\Big(\overline{\boldsymbol{q}}\cdot\boldsymbol{e}_3\Big)+\varepsilon\nabla_{\overline{\boldsymbol{x}}}\cdot\overline{\boldsymbol{q}}=0 & \text{pour } \bar t\in]0,\overline{T}[\,,\ (\overline{\boldsymbol{x}},\bar z)\in\Omega_{\overline{h}}^{+}(\bar t),\\ \alpha\,\overline{P}\left(\varepsilon^2\left\|\nabla'\overline{h}_{\rm soil}\right\|^2+1\right)^{1/2}+\beta\,\overline{\boldsymbol{q}}\cdot\Big(\boldsymbol{e}_3-\varepsilon\nabla'\overline{h}_{\rm soil}\Big)=\left(\varepsilon^2\left\|\nabla'\overline{h}_{\rm soil}\right\|^2+1\right)^{1/2}\overline{F} & \text{sur }]0,\overline{T}[\times\overline{\Gamma}_{\rm soil},\\ \boldsymbol{q}\cdot\boldsymbol{n}=0 & \text{sur }]0,\overline{T}[\times\overline{\Gamma}_{\rm ver}\\ \overline{P}\Big(\bar t,\overline{\boldsymbol{x}},\overline{h}(\bar t,\overline{\boldsymbol{x}})\Big)=\rho g\Big(\overline{H}(\bar t,\overline{\boldsymbol{x}})-\overline{h}(\bar t,\overline{\boldsymbol{x}})\Big) & \text{pour } (\bar t,\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x,\\ \overline{P}(0,\overline{\boldsymbol{x}},\bar z)=\overline{P_{\rm init}}(\overline{\boldsymbol{x}},\bar z) & \text{pour } (\overline{\boldsymbol{x}},\bar z)\in\Omega_{\overline{h}}^{+}(0), \end{cases} \tag{1.135}$$

L'équation satisfaite par la pression dans la nappe phréatique devient

$$\overline{P}(\bar t,\overline{\boldsymbol{x}},\bar z)=\rho g\Big(\overline{H}(\bar t,\overline{\boldsymbol{x}})-\bar z\Big)\quad\text{pour } \bar t\in[0,\overline{T}[\,,\ (\overline{\boldsymbol{x}},\bar z)\in\Omega_{\overline{h}}^{-}(\bar t), \tag{1.136}$$

et celle satisfaite par la charge hydraulique est donnée par

$$\begin{cases} \rho g\,\alpha_p(\overline{h}-\overline{h}_{\rm bot})\dfrac{\overline{T}}{T}\partial_{\bar t}\overline{H}-\varepsilon^2\nabla'\cdot\Big(\overline{\boldsymbol{J}}(\overline{H})\,\nabla'\overline{H}\Big)=-\Big(\overline{\boldsymbol{q}}\cdot(\boldsymbol{e}_3-\varepsilon\nabla\overline{h})\Big)\Big|_{\Gamma_h^+} & \text{pour } (\bar t,\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x,\\ \overline{\boldsymbol{J}}(\overline{H})\nabla'\overline{H}\cdot\overline{\boldsymbol{n}}=0 & \text{pour } (\bar t,\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\partial\overline{\Omega}_x,\\ \overline{H}(0,\overline{\boldsymbol{x}})=\overline{H_{\rm init}}(\overline{\boldsymbol{x}}) & \text{pour } \overline{\boldsymbol{x}}\in\overline{\Omega}_x, \end{cases} \tag{1.137}$$

où la première équation de (1.137) s'écrit encore

$$\begin{aligned} -\varepsilon^2\nabla'\cdot\Big(\overline{\boldsymbol{J}}(\overline{H})\,\nabla'\overline{H}\Big)&=-\Big(\overline{\boldsymbol{q}}\cdot(\boldsymbol{e}_3-\varepsilon\nabla\overline{h}_{\rm soil})\Big)\Big|_{\overline{\Gamma}_{\rm soil}}\\ &-\frac{\overline{T}}{T}\left(\int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h_{\rm soil}}(\overline{\boldsymbol{x}})}\phi\,\partial_{\bar t}s(\overline{P})+\alpha_P\,s(\overline{P})\,\partial_{\bar t}\overline{P}\,d\bar z\right)-\varepsilon\nabla'\cdot\left(\int_{\overline{h}(\bar t,\overline{\boldsymbol{x}})}^{\overline{h_{\rm soil}}(\overline{\boldsymbol{x}})}\overline{\boldsymbol{q}}\right)\quad\text{dans } \in]0,\overline{T}[\times\overline{\Omega}_x, \end{aligned} \tag{1.138}$$

La définition de l'interface séparant les deux types différents d'écoulements est donnée par

$$\overline{h}(\bar t,\overline{\boldsymbol{x}})=\max\left\{\min\left\{\overline{H}(\bar t,\overline{\boldsymbol{x}})-\frac{P_s}{\rho g},\overline{h}_{\max}(\overline{\boldsymbol{x}})\right\},\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})\right\}\qquad\text{pour } (\bar t,\overline{\boldsymbol{x}})\in[0,\overline{T}[\times\overline{\Omega}_x, \tag{1.139}$$

et finalement la vitesse est telle que

$$\begin{cases} \overline{\boldsymbol{v}}=\overline{\boldsymbol{q}}+\overline{\boldsymbol{w}} & \text{pour } \bar t\in]0,\overline{T}[\,,\ (\overline{\boldsymbol{x}},\bar z)\in\overline{\Omega},\\ \overline{\boldsymbol{q}}=-\kappa(\overline{P})\,\overline{K}_{zz}\left(\dfrac{1}{\rho g}\partial_{\bar z}\overline{P}+1\right)\boldsymbol{e}_3-\varepsilon\dfrac{\kappa(\overline{P})}{\rho g}\overline{\boldsymbol{G}}_0\nabla\overline{P}, & \text{pour } \bar t\in]0,\overline{T}[\,,\ (\overline{\boldsymbol{x}},\bar z)\in\overline{\Omega}.\\ \overline{\boldsymbol{w}}=-\varepsilon\,\kappa\Big(\rho g(\overline{H}-\bar z)\Big)\,\overline{\boldsymbol{A}}_0\nabla\overline{H} & \text{pour } \bar t\in]0,\overline{T}[\,,\ (\overline{\boldsymbol{x}},\bar z)\in\overline{\Omega}, \end{cases} \tag{1.140}$$

où

$$\boldsymbol{G}_0=\begin{pmatrix}\boldsymbol{N}_0 & 0\\ 0 & 0\end{pmatrix}.$$

1.4.5.2 Problème effectifs

Notre principal but est de montrer que le modèle classique de Richards (1.116) ainsi que chaque modèle de la classe (1.127)–(1.133) présentent exactement les mêmes comportements dominants pour toutes les échelles de temps considérées (courte si $T=\bar T$, intermédiaire si $T=\bar T/\varepsilon$ et longue si $T=\bar T/\varepsilon^2$). Ce résultat est montré dans le chapitre 4 sous certaines hypothèses et est énoncé dans la proposition 3. De plus, les comportements dominants sont représentés par les problèmes effectifs suivant une analyse du comportement asymptotique du flux, c'est à dire des problèmes (1.134) et (1.135)-(1.140).

Développements asymptotiques formels

- Pour la pression et la vitesse :

$$\overline{P}_\varepsilon^\gamma = \overline{P}_0^\gamma + \varepsilon\,\overline{P}_1^\gamma + \varepsilon^2\,\overline{P}_2^\gamma + \dots \qquad \overline{\boldsymbol{v}}_\varepsilon^\gamma = \overline{\boldsymbol{v}}_0^\gamma + \varepsilon\,\overline{\boldsymbol{v}}_1^\gamma + \varepsilon^2\,\overline{\boldsymbol{v}}_2^\gamma + \dots . \tag{1.141}$$

- Pour les inconnus auxiliaires dans (1.127)–(1.133) :

$$\begin{array}{ll} \overline{\boldsymbol{q}}_\varepsilon^\gamma = \overline{\boldsymbol{q}}_0^\gamma + \varepsilon\,\overline{\boldsymbol{q}}_1^\gamma + \varepsilon^2\,\overline{\boldsymbol{q}}_2^\gamma + \dots & \overline{\boldsymbol{w}}_\varepsilon^\gamma = \overline{\boldsymbol{w}}_0^\gamma + \varepsilon\,\overline{\boldsymbol{w}}_1^\gamma + \varepsilon^2\,\overline{\boldsymbol{w}}_2^\gamma + \dots \\ \overline{H}_\varepsilon^\gamma = \overline{H}_0 + \varepsilon\,\overline{H}_1^\gamma + \varepsilon^2\,\overline{H}_2^\gamma + \dots & \overline{h}_\varepsilon^\gamma = \overline{h}_0^\gamma + \varepsilon\,\overline{h}_1^\gamma + \varepsilon^2\,\overline{h}_2^\gamma + \dots, \end{array} \tag{1.142}$$

  Aucune mise à l'échelle arbitraire n'est imposée ; en particulier, nous ne supposons pas, comme dans [66] que la vitesse verticale est significativement plus petite que la vitesse horizontale lorsque le *rapport* ε est petit.

- pour les termes source :

$$\overline{F}_\varepsilon = \overline{F}_0 + \varepsilon\,\overline{F}_1 + \varepsilon^2\,\overline{F}_2 + \dots . \tag{1.143}$$

- De plus, comme θ et κ sont de classe $\mathcal{C}^\infty$ par morceaux, on écrit :

$$\begin{aligned} \theta(\overline{P}_\varepsilon^\gamma) &= \theta(\overline{P}_0^\gamma) + \varepsilon(\overline{P}_1^\gamma + \varepsilon\,\overline{P}_2^\gamma + \dots)\theta'(\overline{P}_0^\gamma) + \frac{\varepsilon^2}{2}(\overline{P}_1^\gamma + \varepsilon\,\overline{P}_2^\gamma + \dots)^2\theta''(\overline{P}_0^\gamma) + \dots \\ \kappa(\overline{P}_\varepsilon^\gamma) &= \kappa(\overline{P}_0^\gamma) + \varepsilon(\overline{P}_1^\gamma + \varepsilon\,\overline{P}_2^\gamma + \dots)\kappa'(\overline{P}_0^\gamma) + \frac{\varepsilon^2}{2}(\overline{P}_1^\gamma + \varepsilon\,P_2^\gamma + \dots)^2\kappa''(\overline{P}_0^\gamma) + \dots \end{aligned} \tag{1.144}$$

Problèmes effectifs à l'ordre principal :

- Lié à l'échelle en temps court ($T = \overline{T}$),

$$\begin{cases} \phi\,\partial_{\overline{t}} s(\overline{P}_0) - \partial_{\overline{z}}\overline{u}_0 = 0 & \text{dans }]0,\overline{T}[\times\Omega \\ \overline{u}_0 = -\kappa(\overline{P}_0)\,\overline{K}_{zz}\Big(\dfrac{1}{\rho g}\partial_{\overline{z}}\overline{P}_0 + 1\Big) & \text{dans }]0,\overline{T}[\times\Omega \\ \alpha\,\overline{P}_0 + \beta\,\overline{u}_0 = \overline{F}_0 & \text{sur }]0,\overline{T}[\times\overline{\Gamma}_{\text{soil}} \\ \overline{u}_0 = 0 & \text{sur}]0,\overline{T}[\times\overline{\Gamma}_{\text{bot}} \end{cases} \tag{1.145}$$

- Lié à des échelles en temps non court ($T = \varepsilon^{-1}\overline{T}$ ou $T = \varepsilon^{-2}\overline{T}$),

$$\begin{cases} \overline{P}_0(t,\boldsymbol{x},z) = \rho g\Big(\overline{H}_0(t,\boldsymbol{x}) - \overline{z}\Big) & \text{dans }]0,\overline{T}[\times\overline{\Omega} \\ \overline{\boldsymbol{v}}_0 = 0 & \text{dans }]0,\overline{T}[\times\overline{\Omega} \end{cases} \tag{1.146}$$

- Lié à des échelles en temps non courts ($T = \varepsilon^{-1}\overline{T}$ ou $T = \varepsilon^{-2}\overline{T}$) si $\alpha \neq 0$,

$$\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) = \frac{\overline{F}_0(\overline{t},\overline{\boldsymbol{x}})}{\alpha\,\rho\,g} + \overline{h}_{\text{soil}}(\overline{t},\overline{\boldsymbol{x}}) \qquad \text{dans }]0,\overline{T}[\times\overline{\Omega}_x \tag{1.147}$$

- Lié à l'échelle en temps intermédiaire ($T = \varepsilon^{-1}\overline{T}$) si $\alpha = 0$ (et donc $\beta \neq 0$),

$$\rho g\Big(\int_{\overline{h}_{\text{bot}}}^{\overline{h}_{\text{soil}}} \phi\,s'(\overline{P}_0) + \alpha_P\,s(\overline{P}_0)\,d\overline{z}\Big)\partial_t\overline{H}_0 = -\frac{\overline{F}_1}{\beta} \qquad \text{dans }]0,\overline{T}[\times\overline{\Omega}_x \tag{1.148}$$

- Lié à l'échelle en temps long ($T = \varepsilon^{-2}\overline{T}$) si $\alpha = 0$ (et donc $\beta \neq 0$),

$$\begin{cases} \displaystyle\int_{\overline{h}_{\text{bot}}}^{\overline{h}_{\text{soil}}} \phi\,\partial_{\overline{t}} s(\overline{P}_0) + \alpha_P\,s(\overline{P}_0)\partial_t\overline{P}_0\,d\overline{z} - \nabla'\cdot\Big(\overline{\boldsymbol{K}}(\overline{H}_0)\,\nabla'\overline{H}_0\Big) = -\dfrac{\overline{F}_2}{\beta} & \text{dans }]0,\overline{T}[\times\overline{\Omega}_x \\ \overline{\boldsymbol{K}}(\overline{H}_0)\,\nabla_{\overline{x}}\overline{H}_0\cdot\overline{\boldsymbol{n}} = 0 & \text{sur }]0,\overline{T}[\times\overline{\Gamma}_{\text{ver}} \end{cases} \tag{1.149}$$

  et concernant le premier ordre de la vitesse

$$\overline{\boldsymbol{v}}_1 = -\overline{\kappa}(\overline{P}_0)\,\overline{\boldsymbol{M}}_0\,\nabla_{\overline{x}}\overline{H}_0 \qquad \text{dans }]0,\overline{T}[\times\overline{\Omega} \tag{1.150}$$

Proposition 3 *Soit $(\overline{P}_\varepsilon^\gamma, \overline{\boldsymbol{v}}_\varepsilon^\gamma)$ la solution du problème de Richards $3D$ adimensionné (1.134) ou du modèle couplé adimensionné (1.135)–(1.139) pour $T = \varepsilon^{-\gamma}\overline{T}$ et $\gamma \in \{0,1,2\}$. On suppose que (1.141)–(1.144) est vrai. Les termes d'ordre principal de la pression et de la vitesse du fluide sont caractérisés par :*

(i) $\overline{P}_0^0$ *satisfait (1.145) sous l'hypothèse supplémentaire* $\alpha_p = 0$.

(ii) $(\overline{P}_0^1, \overline{\boldsymbol{v}}_0^1)$ *satisfait (1.146) et (1.147) si* $\alpha \neq 0$, *ou (1.146) et (1.148)avec la condition de compatibilité* $\overline{F}_0 = 0$ *si* $\alpha = 0$.

(iii) $(\overline{P}_0^2, \overline{\boldsymbol{v}}_0^2)$ *satisfait (1.146) et (1.147) si* $\alpha \neq 0$, *ou (1.146) et (1.149) avec la condition de compatibilité* $\overline{F}_0 = \overline{F}_1 = 0$ *si* $\alpha = 0$. *De plus, le terme d'ordre suivant de la vitesse* $\overline{\boldsymbol{v}}_1^2$ *satisfait (1.150) si* $\alpha = 0$.

L'idée du preuve consiste à substituer les développements asymptotiques formels dans le problème adimensionné considéré ((1.134) ou (1.135)– (1.139)). Une cascade d'équations s'ensuit en identifiant les puissances de ε. Ensuite, nous caractérisons formellement les termes dominants de ces développement et nous les obtenons ainsi comme solutions des problèmes effectifs (1.145)–(1.149). On arrive donc à justifier formellement que le modèle couplé (1.127)–(1.133) approche bien (asymptotiquement) le problème original de Richards $3D$ (1.116) dans des aquifères peu profonds quelque soit l'échelle de temps considérée.

Conclusions Nous concluons ce paragraphe par quelques commentaires concernant les conséquences de la proposition 3. À ce stade on a :

- quelque soit le choix de $\boldsymbol{G}_0$, le problème (1.127)–(1.133) est une bonne approximation du problème $3D$ de Richards compressible (1.117) pour décrire l'écoulement dans des aquifères peu profonds :
  - à l'échelle des temps court si $\alpha_p = 0$
  - à l'échelle des temps intermédiaire et long, quelque soit le choix de α_p.
- le problème (1.127)–(1.133) a la même structure que l'original (1.120)–(1.125). Il peut être vu également comme une petite perturbation de (1.120)–(1.125) lorsqu'on considère un petit tenseur $\boldsymbol{G}_0$ et un petit paramètre de compressibilité α_P.
- par contruction, l'équation dans la partie supérieure de l'aquifére est maintenant un problème de Richards $3D$ compressible. Cela augmente les contrôles à priori sur la solution et élimine la dégénérescence, ce qui rend le problème (1.127)–(1.133) moins difficile à étudier théoriquement que (1.120)–(1.125).

1.4.5.3 Reformulation du problème

On suppose que le niveau h caractérisant le niveau sous lequel l'écoulement vertical est supposé instantané, reste loin de h_{bot}. On suppose également que ce niveau n'atteint pas h_{soil}. Plus précisément, on supose qu'il existe $\delta > 0$ tel que :

$$h_{\text{bot}} + \delta \leq h(t,x) < h_{\text{soil}} \quad \forall (t,x) \in (0,T) \times \Omega.$$

Puis d'après (1.133), on obtient : $h = \tilde{H} - \frac{P_s}{\rho g}$.
Dès cette partie, nous allégerons les notations en supprimant l'indice "0" dans les tenseurs de conductivité et en n'utilisant plus les caractères gras pour la représentation des vecteurs et des matrices. En particulier :

$$B = \boldsymbol{B}_0, \quad S = \boldsymbol{S}_0, \quad \tilde{J} = \tilde{\boldsymbol{J}}.$$

Compte tenu des notations ci-dessus, le modèle final $(\mathcal{N})$ couplant l'écoulement $3D$-Richards et l'écoulement horizontal de Dupuit consiste en les systèmes (1.151) et (1.153) :

- Dans Ω_t l'équation $3D$-Richards suivante est valable :

$$\begin{cases} \partial_t\theta(P)+\theta\alpha_P\,\partial_t P+\nabla\cdot q=0 & \text{dans} \quad (0,T)\times\Omega_t,\\ a\,P+q\cdot\vec{\nu}=F & \text{sur} \quad (0,T)\times\Gamma_{\text{soil}},\\ q\cdot\vec{\nu}=0 & \text{sur} \quad (0,T)\times\Gamma_{\text{ver}},\\ P\Big(t,x,h(t,x)\Big)=P_s & \text{dans} \quad (0,T)\times\Omega_x,\\ P(0,x,z)=P_0(x,z) & \text{dans} \quad \Omega_0. \end{cases} \tag{1.151}$$

  La vitesse effective q est donnée par :

$$q=-\kappa(P)B\nabla\left(\frac{P}{\rho\,g}+z\right). \tag{1.152}$$

- Dans Ω_t^- la pression P satisfait :

$$P(t,x,z)=\rho\,g\left(\frac{P_s}{\rho g}+h-z\right) \quad \text{dans } (0,T)\times\Omega_t^-.$$

- La profondeur de Γ_t, h, satisfait dans Ω_x

$$\begin{cases} S_0B_f\partial_t h-\nabla'\cdot(\tilde{J}\nabla' h)=-\displaystyle\int_{h(t,x)}^{h_{\text{soil}}(x)}\Big(\partial_t\theta(P)+\theta(P)\alpha_P\partial_t P\Big)\,dz\\ \qquad\qquad -q_{|z=h^+_{\text{soil}}}\cdot\vec{\nu}-\nabla'\cdot\Big(\int_h^{h_{\text{soil}}}q\,dz\Big),\\ \tilde{J}\nabla' h\cdot\vec{\nu}=0 \qquad \text{sur } (0,T)\times\partial\Omega_x,\\ h(0,x)=h_0(x) \quad \text{dans } \Omega_x. \end{cases} \tag{1.153}$$

La condition de Neumann homogène sur $\partial\Omega_x$ est supposée pour simplifier la présentation.

Difficultés pour l'étude théorique d'un tel modèle Les principales difficultés inhérentes au modèle (1.120)–(1.125) sont les suivantes : l'intégration des équations sur un domaine à frontière libre, la nonlinéarité et la dégénérescence apparaissant dans les dérivées en temps des deux équations et la perte de contrôle des composantes horizontales de la pression. Toutes ces difficultés étaient déjà présentes dans l'étude du modèle décrit dans la section 1.4.2 (modèle $\mathcal{M}$) à l'exception de la dégénérescence en temps et en espace présentes dans la seconde équation du modèle ($\mathcal{N}$), l'intégration verticale de la conductivité hydraulique permettant de neutraliser ces difficultés pour le modèle ($\mathcal{M}$). Nous considérons maintenant une forme très générale de conductivité hydraulique (comme cela a été fait dans [48]), ce qui induit une nonlinéarité supplémentaire dans la dérivée temporelle de l'équation en h. Notons également que, contrairement au modèle ($\mathcal{M}$), nous considérons une condition aux limites de Robin très générale à l'interface Γ_{soil} permettant ainsi de décrire les échanges entre les eaux de surface et les eaux souterraines.
L'analyse des équations de Richards est connue pour être délicate à cause de la présence des coefficients dégénérés apparaissant dans les termes diffusifs et dans les dérivées temporelles. Habituellement, la transformée de Kirchov est utilisée pour éliminer la nonlinéarité dans le terme diffusif. Mais, ainsi que nous l'avons dit, la prise en compte de la faible compressibilité de l'eau induira des transformations bijectives qui auront le même impact sur les dégénérescences et nonlinéarités des dérivées en temps des équations des systèmes ($\mathcal{N}$) et ($\mathcal{M}$) que la transformée de Kirchov sur le terme diffusif, c'est donc pour ce choix de transformations que nous optons.
Concernant, la dépendance en temps du domaine d'intégration, il existe plusieurs méthodes pour aborder l'étude des problèmes à frontière libre. Nous choisissons ici le cadre de travail des domaines non cylindriques introduit par Lions et Mignot qui consiste à étendre la solution par zéro en dehors

du domaine variable ramenant ainsi l'étude du problème à celle d'un problème dans un domaine fixe. Finalement, la première équation (1.120) paraît être mal posée d'un point de vue théorique. En effet, la conductivité hydraulique horizontale étant nulle, il s'agit d'une équation $1D$ mais définie dans un domaine $3D$, d'où un perte de contrôle du gradient de la pression par rapport aux variables horizontales. Le fait de considérer une faible conductivité hydraulique horizontale dans la zone insaturée de l'aquifère permet de pallier à cette perte d'information sur les composantes horizontales de P.

1.4.6 Analyse mathématique des modèles couplés

L'étude mathématique des modèles $(\mathcal{M})$ et $(\mathcal{N})$ est particulièrement délicate en raison des non-linéarités, de la frontière libre entre chaque zone et de la difficulté résultant du couplage entre les deux zones qui s'exprime ici en terme de flux à l'interface. Nous devons également faire face aux difficultés mathématiques inhérentes aux équations de Richards. De plus, il existe une difficulté mathématique générale dans la structure de l'ensemble des PDE modélisant la dynamique des eaux souterraines. En effet, lorsqu'on considère une nappe phréatique libre, il faut faire face à la disparition progressive de l'eau dans la zone de désaturation et donc à la disparition d'une des principales inconnues du problème.

Il existe une littérature abondante concernant les équations classiques de Richards. Mentionnons les travaux incontournables de Alt *et al* [50, 51] ainsi que les articles [52, 53, 55] dédiés à l'étude de l'équation "dégénérée" en temps :

$$\partial_t \theta(p) - \nabla p = 0,$$

où $\theta(p)$ désigne la teneur volumétrique en humidité. Citons également dans le cas unidimensionnel le travail de Yin [56] concernant l'existence d'une solution faible pour le problème totalement dégénéré :

$$\partial_t \theta(p) - \partial_x(\kappa(\theta(p))\partial_x p) = 0,$$

en supposant tout simplement que $\theta', \kappa' > 0$.
Classiquement, la transformée de Kirchoff est appliquée à l'équation de Richards (sous des hypothèses appropriées sur la porosité et la perméabilité) pour éliminer la non-linéarité dans le terme diffusif. Dans ce travail, nous exploitons l'hypothèse de la faible compressibilité de l'eau pour éliminer la dégénérescence apparaissant dans la dérivée temporelle de l'équation de Richards. Cette transformation nous ramène au cadre des équations paraboliques quasi-linéaires sur des domaines non cylindriques auxquelles on peut appliquer la méthode des domaines auxiliaires introduite par Lions et Mignot [57, 58] pour traiter le problème à frontière libre (1.112) et (1.151) . Par ailleurs, pour le premier modèle décrit dans ce travail (modèle $(\mathcal{M})$, section 1.4.2), le moyenne verticale dans la zone de saturation conduit à une équation elliptique dégénérée dont la dégénérescence dépend de l'épaisseur de la zone saturée. La prise en compte de la compressibilité de l'eau introduit une dégénérescence dans la dérivée temporelle dépendant également de l'épaisseur de la zone saturée. Un changement de variable permet alors d'absorber les deux termes dégénérés et de revenir à une équation parabolique régulière. Pour le modèle $(\mathcal{N})$ décrit dans 1.4.3, la forme très générale de la conductivité hydraulique (1.132) que nous considérons induit une non linéarité supplémentaire dans la dérivée en temps. Cela change donc totalement l'étude de la seconde équation régissant l'évolution de l'interface (équation en h). Dans ce cas, on ne peut absolument plus faire le changement de variable qui permettait d'absorber la non linéarité en temps dans le cas du modèle $(\mathcal{M})$. Deux analyses mathématiques des modèles de type Dupuit-Richards sont donc présentées dans la suite du document. Elles diffèrent selon qu'on considère le cas isotrope et non conservatif (modèle $(\mathcal{M})$) ou le cas anisotrope et conservatif (modèle $(\mathcal{N})$).

Tout d'abord, il est nécessaire d'introduire quelques notations et rappeler des résultats généraux mathématiques utiles pour la suite de l'étude de ces deux modèles.

1.4.6.1 Notations et résultats auxiliaires

Soit $\mathcal{O}$ le domaine ouvert de $\mathbb{R}\times\mathbb{R}^N$, inclus dans l'ensemble $\mathbb{R}^+\times\Omega$ et défini par :

$$\mathcal{O}=\mathbb{R}^+\times\Omega_x\times(h,h_{\text{soil}}),$$

où h désigne la position de l'interface Γ_t. On pose :

$\Omega_{t'}=\mathcal{O}\cap\{t=t'\},\quad\forall t'\geq 0,\quad$ (définition compatible avec (1.85))
$\Omega'_{t'}=\Omega\setminus\Omega_{t'},\quad\forall t'\geq 0,$
$\mathcal{O}_T=\mathcal{O}\cap\{0\leq t\leq T\},$
$\mathcal{O}'_T=((0,T)\times\Omega)\setminus\mathcal{O}_T,$
$\Gamma'=\Gamma\setminus\Omega_0,\quad$ (la frontière latérale de $\mathcal{O}$),
$\gamma_t=\partial\Omega_t\quad$ (la frontière de $\Omega_t\subset\mathbb{R}^N$),
$\Gamma'_T=\Gamma'\cap\{0<t<T\}=\cup_{t\in(0,T)}\gamma_t.$

On définit :

$$H^{0,1}(\mathcal{O})=\Big\{u\mid D^pu\in L^2(\mathcal{O})\ \text{ pour }\ |p|\leq 1\Big\}\quad,\quad D^pu=\Big\{D^\alpha u\mid\alpha=(\alpha_1,\alpha_2,\alpha_3)\ \text{ avec }\ |\alpha|=p\Big\}.$$

C'est un espace de Hilbert muni de la norme :

$$\|u\|_{H^{0,1}(\mathcal{O})}=\Big(\Sigma_{p\leq 1}\int_{\mathcal{O}}|D^pu|\,dx\,dt\Big)^{1/2}.$$

$H_0^{0,1}(\mathcal{O})$ désigne la fermeture de $\mathcal{D}(\mathcal{O})$ dans $H^{0,1}(\mathcal{O})$ pour la norme $\|.\|_{H^{0,1}(\mathcal{O})}$. Dans ce qui suit, on notera $F(\mathcal{O})=H_0^{0,1}(\mathcal{O})$ et $F'(\mathcal{O})$ son dual topologique. En outre, on introduit les espaces :

$$\mathcal{A}(\mathcal{O})=\Big\{u\mid u\in H^{0,1}(\mathcal{O})\,,\,\frac{du}{dt}\in F'(\mathcal{O})\Big\}\quad\text{et}\quad\mathcal{B}(\mathcal{O})=\Big\{u\mid u\in F(\mathcal{O})\,,\,\frac{du}{dt}\in F'(\mathcal{O})\Big\},$$

munis des normes Hilbertiennes :

$$\|\cdot\|_{\mathcal{A}(\mathcal{O})}=\Big(\|\cdot\|^2_{H^{0,1}(\mathcal{O})}+\|\partial_t\cdot\|^2_{(H^{0,1}(\mathcal{O}))'}\Big)^{1/2}\quad\text{et}\quad\|\cdot\|_{\mathcal{B}(\mathcal{O})}=\Big(\|\cdot\|^2_{H_0^{0,1}(\mathcal{O})}+\|\partial_t\cdot\|^2_{F'(\mathcal{O})}\Big)^{1/2}.$$

Finallement, $\mathcal{B}_0(\mathcal{O})$ (resp. $\mathcal{B}_T(\mathcal{O})$) désigne la fermeture dans $\mathcal{B}(\mathcal{O})$ des fonctions nulles au voisinage de $t=0$ (resp. $t=T$).

Nous énonçons maintenant quelques résultats auxiliaires prouvés dans [57].

Lemme 1 *Si $\mathcal{O}$ est suffisamment régulier, alors on a :*

1. $H^{0,1}(\mathcal{O}_T)=L^2([0,T];H^1(\Omega_t))$ *où*

$$L^2([0,T];H^1(\Omega_t))=\Big\{u\mid u(t,.)\in H^1(\Omega_t))\,,\,t\in[0,T]\ p.p\ \textit{et}\ \|u\|_{H^{0,1}(\mathcal{O}_T)}<+\infty\Big\},$$

avec $\|u\|_{H^{0,1}(\mathcal{O}_T)}=\int_0^T\|u\|^2_{H^1(\Omega_t)}\,dt$.
Un résultat similaire vaut pour $H_0^{0,1}(\mathcal{O}_T)$.

2. *Pour* $u\in H_0^{0,1}(\mathcal{O})$, *on peut définir* $\gamma(u)$, *la trace de* u *sur* Γ' *dans* $L^2(\Gamma')$. *De plus,*

$$u\in F(\mathcal{O})\Longleftrightarrow\gamma(u)=0.$$

3. *Soit* $u\in\mathcal{B}(\mathcal{O}_T)$, *alors* $u\in\mathcal{B}_T(\mathcal{O})\Longleftrightarrow u(T,.)=0.$

4. $\forall u, v \in \mathcal{O}_s$), *on a :*

$$\langle \frac{\partial u}{\partial t}, v \rangle_{F',F} + \langle u, \frac{\partial v}{\partial t} \rangle_{F',F} = (u(s,.), v(s,.))_{L^2(\Omega_s)} - (u(0,.), v(0,.))_{L^2(\Omega_0)}. \tag{1.154}$$

Soit Ω' un ouvert borné de $\mathbb{R}^3$. Pour des raisons de brièveté, nous noterons $H^1(\Omega') = W^{1,2}(\Omega')$ et

$$V(\Omega') = H_0^1(\Omega'), \quad V'(\Omega') = H^{-1}(\Omega'), \quad H(\Omega') = L^2(\Omega').$$

Nous rappelons que les injections $V(\Omega') \subset H(\Omega') = H'(\Omega') \subset V'(\Omega')$ sont denses et compactes. Pour tout $T > 0$, soit $W(0,T,\Omega')$ désignant l'espace

$$W(0,T,\Omega') := \left\{ \omega \in L^2(0,T;V(\Omega')),\ \partial_t \omega \in L^2(0,T;V'(\Omega')) \right\}$$

muni de la norme Hilbertienne $\| \cdot \|_{W(0,T,\Omega')} = \left(\| \cdot \|^2_{L^2(0,T;V(\Omega'))} + \| \partial_t \cdot \|^2_{L^2(0,T;V'(\Omega'))} \right)^{1/2}$. Les injections suivantes sont continues ([60] prop. 2.1 et thm. 3.1, chapitre 1)

$$W(0,T,\Omega') \subset \mathcal{C}([0,T]; [V(\Omega'), V'(\Omega')]_{\frac{1}{2}}) = \mathcal{C}([0,T]; H(\Omega')),$$

tandis que l'injection

$$W(0,T,\Omega') \subset L^2(0,T;H(\Omega')) \tag{1.155}$$

est compacte (lemme d'Aubin, voir [61]).
Le résultat suivant dû à F. Mignot (voir [62]) est utilisé dans la suite.

Lemme 2 *Soit $f : R \to R$ une application croissante et continue telle que*

$$\limsup_{|\lambda| \to +\infty} |f(\lambda)/\lambda| < +\infty.$$

Soit $\omega \in L^2(0,T;H(\Omega'))$ tel que $\partial_t \omega \in L^2(0,T;V'(\Omega'))$ et $f(\omega) \in L^2(0,T;V(\Omega'))$. Alors,

$$\langle \partial_t \omega, f(\omega) \rangle_{(V(\Omega'))',V(\Omega')} = \frac{d}{dt} \int_\Omega \left(\int_0^{\omega(\cdot,y)} f(r)\, dr \right) dy \ \text{ dans } \ \mathcal{D}'(0,T).$$

Ainsi, pour tout $0 \leq t_1 < t_2 \leq T$

$$\int_{t_1}^{t_2} \langle \partial_t \omega, f(\omega) \rangle_{V',V}\, dt = \int_\Omega \left(\int_{\omega(t_1,y)}^{\omega(t_2,y)} f(r)\, dr \right) dy.$$

Remarque 2 *le résultat (1.154) du Lemme 1 est une généralisation du Lemme 2 au cas où le domaine spatial Ω' dépend du temps.*

1.4.6.2 Résultats principaux

Nous cherchons à donner un résultat d'existence de solutions faibles physiquement admissibles pour les modèles $(\mathcal{M})$ et $(\mathcal{N})$ complétés par des conditions initiales et aux limites.
Nous commençons par les caractéristiques de la structure poreuse.

- L'étude du modèle $(\mathcal{M})$ se limite au cas isotropique ainsi le tenseur K_0 est supposé être un scalaire. Dans la partie saturée de l'aquifère, la conductivité hydraulique moyenne $\tilde{K}$ est donc égale à la constante $\frac{K_0 \rho_0 g}{\mu}$. Sans perte de généralité, nous supposerons un terme source nul dans l'equation (1.110) (c'est à dire $\tilde{Q} = 0$).

- En revanche, on considère le cas anisotropique pour l'étude du modèle $(\mathcal{N})$ ainsi la conductivité hydraulique prend une forme très générale. Dans ce cas on suppose que les profondeurs h_{bot} et h_{soil} sont constantes et telles que $h_{\text{bot}} > h_{\text{soil}} > 0$.
  Rappelons que :
  $$q = -\kappa(P)\, B\nabla\Big(\frac{P}{\rho g} + z\Big),$$
  et le tenseur de la conductivité moyennée $\tilde{J}$ de taille 2×2 est défini dans $(0,T)\times\Omega_x$ pour toute fonction $\tilde{H} = \tilde{H}(t,x)$ par :
  $$\tilde{J}(\tilde{H})(t,x) = \int_{h_{\text{bot}}}^{h(x)} S(x,z)\, dz + \int_{h(x)}^{h_{\text{soil}}} \kappa\Big(\rho g(\tilde{H}(t,x) - z)\Big)\,(S(x,z) - N)\, dz. \tag{1.156}$$

Les fonctions θ et κ dépendent de la pression. Nous supposons que :

$$\theta \in \mathcal{C}^1(\mathbb{R}), \qquad 0 < \theta_- := \phi\, s_0 \le \theta(x) \le \theta_+, \qquad \theta'(x) \ge 0 \quad \forall x \in \mathbb{R}. \tag{1.157}$$

$$\kappa \in \mathcal{C}(\mathbb{R}), \qquad 0 < \kappa_- \le \kappa(x) \le \kappa_+ \quad \forall x \in \mathbb{R}. \tag{1.158}$$

Avant d'énoncer le résultat principal de ce travail, nous allons transformer le problème d'origine et nous ramener au cadre introduit dans [58].
Les hypothèses ci-dessus sur le fluide et le milieu permettent d'éliminer la non-linéarité en temps des équations (1.102) et (1.151), à savoir les hypothèses (1.157) et (1.158) sont suffisantes pour définir la fonction primitive $\mathcal{P}$ telle que :

$$\mathcal{P}(P) = \theta(P) + \alpha_P \int^P \theta(s)\, ds.$$

Un calcul direct donne $\mathcal{P}'(P) = \theta'(P) + \alpha_P\,\theta(P) > \alpha_P\,\theta_- > 0$, en effet d'après l'hypothèse précédente, on a $\theta'(P) > 0$ et $\theta(P) > \phi\, s_0$.
Comme $\mathcal{P}$ est une application bijective, alors l'existence de p telle que

$$p = \mathcal{P}(P)$$

équivaut à l'existence d'une solution P du problème de Richards original pour les deux modèles couplés $(\mathcal{M})$ et $(\mathcal{N})$.

Les transformées par l'application $\mathcal{P}$ des équations (1.102) et (1.151) sont respectivement données par (1.159) et (1.160) :

$$\partial_t p - \frac{1}{\mu}\nabla\cdot\Big(\frac{\kappa(\mathcal{P}^{-1}(p))}{(\theta' + \alpha_P\,\theta)(\mathcal{P}^{-1}(p))}\, K_0\nabla p\Big) - \frac{\rho_0\, g}{\mu}\nabla\cdot\Big(\kappa(\mathcal{P}^{-1}(p)) K_0\, \vec{e_3}\Big) = Q \tag{1.159}$$

$$\partial_t p - \frac{1}{\rho g}\nabla\cdot\Big(\frac{\kappa(\mathcal{P}^{-1}(p))}{(\theta' + \alpha_P\theta)(\mathcal{P}^{-1}(p))} B\nabla p\Big) - \nabla\cdot\Big(\kappa(\mathcal{P}^{-1}(p)) B\vec{e_3}\Big) = 0. \tag{1.160}$$

Finalement, nous introduisons les notations :

$$\tau(p) = \frac{K_0}{\mu}\frac{\kappa(\mathcal{P}^{-1}(p))}{(\theta' + \alpha_P\,\theta)(\mathcal{P}^{-1}(p))} \quad \text{et} \quad \sigma(p) = \frac{1}{\rho g}\frac{\kappa(\mathcal{P}^{-1}(p))}{(\theta' + \alpha_P\theta)(\mathcal{P}^{-1}(p))}.$$

D'après les hypothèses (1.157) et (1.158), on remarque qu'il existe deux couples de réels positifs (τ_-,τ_+) et (σ_-,σ_+) tels que :

$$0 < \tau_- := \frac{K_0\,\kappa_-}{\mu\,\alpha_P\,\theta_+} \le \tau(p) \le \tau_+ := \frac{K_0\,\kappa_+}{\mu\,\alpha_P\,\theta_-} \tag{1.161}$$

$$0 < \sigma_- := \frac{\kappa_-}{\rho g\,\alpha_P\theta_+} \le \sigma(p) \le \sigma_+ := \frac{\kappa_+}{\rho g\,\alpha_P\theta_-}. \tag{1.162}$$

Par ailleurs, concernant le modèle $(\mathcal{N})$, les tenseurs B et $\tilde{J}$ sont supposés bornés et uniformément elliptiques. Plus précisément, il existe deux couples de nombres réels positifs $0 < K^- \leq K^+$ et $0 < \tilde{K}^- \leq \tilde{K}^+$ tels que :

$$0 < K^-|\xi|^2 \leq B\xi \cdot \xi = \sum_{k,l=1}^{3} K_{kl}\xi_k\xi_l \leq K^+|\xi|^2, \ \forall \xi \in \mathbb{R}^3 \setminus \{0\}, \tag{1.163}$$

$$0 < \tilde{K}^-|\xi|^2 \leq \tilde{J}\xi \cdot \xi = \sum_{k,l=1}^{2} \tilde{K}_{kl}\xi_k\xi_l \leq \tilde{K}^+|\xi|^2, \ \forall \xi \in \mathbb{R}^2 \setminus \{0\}. \tag{1.164}$$

Soient $\delta \in \mathbb{R}$ un nombre positif et $l = (h_{\text{soil}} - h_{\text{bot}})$ la fonction (dépendant de l'espace dans le cas du premier modèle) désignant l'épaisseur totale du sous-sol. Nous introduisons les fonctions T_l et U_l définies par :

$$T_l(u) = \sqrt{u} + h_{\text{bot}}, \quad \forall u \in [\delta^2, l^2] \qquad \text{et} \qquad U_l(u) = h - h_{\text{bot}}, \quad \forall u \in [h_{\text{bot}} + \delta, h_{\text{sol}}]$$

qui sont étendues continûment et de manière constante en dehors des intervalles $[\delta^2, l^2]$ et $[h_{\text{bot}} + \delta, h_{\text{sol}}]$ respectivement. Notons que la fonction T_l (resp. U_l) est définie pour l'étude du modèle $(\mathcal{M})$ (resp. $(\mathcal{N})$).
Pour le modèle $(\mathcal{N})$, l'hypothèse sur le paramètre δ est suffisante pour définir une deuxième fonction primitive $\mathcal{T}$ telle que :

$$\mathcal{T}(h) = \begin{cases} \delta(h - h_{\text{bot}}) - \dfrac{\delta^2}{2} & \text{si} \quad h_{\text{bot}} + \frac{\delta}{2} \leq h \leq h_{\text{bot}} + \delta, \\ \dfrac{(h - h_{\text{bot}})^2}{2} & \text{si} \quad h_{\text{bot}} + \delta \leq h \leq h_{\text{soil}}, \\ l\,(h - h_{bot}) - \dfrac{l^2}{2} & \text{si} \quad h \geq h_{\text{soil}}. \end{cases}$$

Un calcul direct nous donne $\mathcal{T}'(h) = U_l(h) > \delta > 0$. Comme $\mathcal{T}$ est une application bijective, alors l'existence de u tel que $u = \mathcal{T}(h)$ équivaut à l'existence d'une solution h de l'équation originale (1.153), de plus on a :

$$\mathcal{T}^{-1}(u) = \begin{cases} \dfrac{u}{\delta} + \dfrac{\delta}{2} + h_{\text{bot}} & \text{si} \quad 0 \leq u \leq \dfrac{\delta^2}{2}, \\ \sqrt{2\,u} & \text{si} \quad \dfrac{\delta^2}{2} \leq u \leq \dfrac{l^2}{2}, \\ \dfrac{u}{l} + \dfrac{l}{2} + h_{\text{bot}} & \text{si} \quad u \geq \dfrac{l^2}{2}. \end{cases}$$

Par conséquent, on a :

$$\frac{1}{l} \leq \|\left(\mathcal{T}^{-1}\right)'\|_\infty \leq \frac{1}{\delta}. \tag{1.165}$$

Remarque 3 *Le petit paramètre δ est introduit dans l'étude du modèle $(\mathcal{N})$ pour contrôler le terme dégénéré en temps dans l'équation (1.153). L'interprétation physique de cet ajout est qu'il y a une quantité d'eau, d'épaisseur au moins égale à δ, partout dans l'aquifère. De plus, une seconde dégénérescence est également cachée dans ce modèle. En effet, pour étendre la solution p en dehors du domaine dépendant du temps Ω_t, il faut aussi imposer à la fonction h qu'elle soit supérieure ou égale à une quantité strictement supérieure à h_{bot}. C'est pour cela encore qu'on a introduit le petit paramètre δ dans la borne inférieure de l'intervalle d'étude. Pour la même raison, il est nécessaire d'imposer ce paramètre dans l'étude du modèle $(\mathcal{M})$.*

Soit $\chi(u)$ la fonction définie par :

$$\chi(u) = \begin{cases} 0 & \text{if } u \leq 0 \\ 1 & \text{if } u > 0 \end{cases}. \tag{1.166}$$

Afin de garantir la non-négativité de u, le contrôle $\chi(u)$ est ajouté devant le côté droit de (1.167) et (1.168).

- Posons $u = (h - h_{\text{bot}})^2$, l'équation (1.110) devient ainsi :

$$\frac{S_0}{2}\,\partial_t u - \frac{\tilde{K}}{2}\nabla'\cdot(\nabla' u) = -\chi(u)\int_{T_l(u(t,x))}^{h_{soil}(x)}\Big(\frac{\partial p}{\partial t} - Q\Big)\,dz. \tag{1.167}$$

- La transformée $\mathcal{T}$ de l'équation (1.131) est alors

$$S_0\partial_t u - \nabla'\cdot\Big(\tilde{J}(\mathcal{T}^{-1}(u))\nabla'\mathcal{T}^{-1}(u)\Big) = -\chi(u)\Big(\int_{h_{\text{bot}}+U_l(\mathcal{T}^{-1}(u))}^{h_{\text{soil}}}\partial_t p\,dz \tag{1.168}$$

$$+\text{div}_x\Big(\int_{h_{\text{bot}}+U_l(\mathcal{T}^{-1}(u))}^{h_{\text{soil}}} q\,dz\Big) + q_{|z=h_{\text{soil}}^+}\cdot\vec{\nu}\Big) \tag{1.169}$$

Définition 1 *Pour les deux modèles couplés, la définition de la profondeur h est dérivée de la construction de u :*

- *pour u donné par (1.167), on définit :*

$$h(t,x) := T_l(u). \tag{1.170}$$

- *pour u donné par (1.168), on pose :*

$$h(t,x) := h_{\text{bot}} + U_l(\mathcal{T}^{-1}(u)). \tag{1.171}$$

Remarque 4 *Ces définitions de h permettent de définir le domaine d'intégration Ω_t (puis l'interface Γ_t) dans les systèmes (1.172)-(1.173) et (1.176)-(1.177). Nous soulignons que par définition, h reste toujours dans l'intervalle $[h_{bot}+\delta, h_{soil}]$.*

Pour chaque modèle couplé, nous sommes maintenant amenés à considérer le problème complété par les conditions aux limites et initiales.

- pour le modèle $(\mathcal{M})$:

$$\partial_t p - \nabla\cdot\Big(\tau(p)\,\nabla p\Big) - \nabla\cdot\Big(\kappa(\mathcal{P}^{-1}(p))K_0\vec{e_3}\Big) = Q \quad \text{dans } \mathcal{O}_T, \tag{1.172}$$

$$p\Big|_{\Gamma_t} = \mathcal{P}(P_s) \quad \text{dans } (0,T), \qquad \nabla\Big(\mathcal{P}^{-1}(p) + \rho_0 g z\Big)\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times(\Gamma_{\text{soil}}\cup\Gamma_{\text{ver}}),$$

$$p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{dans } \Omega_0, \tag{1.173}$$

$$\frac{S_0}{2}\,\partial_t u - \frac{\tilde{K}}{2}\nabla'\cdot(\nabla' u) = -\chi(u)\int_{T_l(u(t,x))}^{h_{\text{soil}}(x)}\Big(\frac{\partial p}{\partial t} - Q\Big)\,dz \quad \text{dans } (0,T)\times\Omega_x, \tag{1.174}$$

$$\nabla u\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times\partial\Omega_x, \qquad u(0,x) = (h_0(x) - h_{\text{bot}}(x))^2 \quad \text{dans } \Omega_x. \tag{1.175}$$

- pour le modèle $(\mathcal{N})$:

$$\partial_t p - \nabla\cdot\Big(\sigma(p)\,B\,\nabla p\Big) - \nabla\cdot\Big(\kappa(\mathcal{P}^{-1}(p))B\vec{e_3}\Big) = 0 \quad \text{dans } \mathcal{O}_T, \tag{1.176}$$

$$p\Big|_{\Gamma_t} = \mathcal{P}(P_s) \quad \text{dans } (0,T), \qquad \nabla\Big(\mathcal{P}^{-1}(p) + \rho g z\Big)\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times\Gamma_{ver},$$

$$a\,P + \nabla\Big(\mathcal{P}^{-1}(p) + b\rho_0 g z\Big)\cdot\vec{\nu} = F \quad \text{sur} \quad (0,T)\times\Gamma_{\text{soil}}, \qquad p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{dans } \Omega_0. \tag{1.177}$$

$$S_0\partial_t u - \nabla'\cdot\Big(\tilde{J}(\mathcal{T}^{-1}(u))\nabla'\mathcal{T}^{-1}(u)\Big) = -\chi(u)\Big(\int_{h_{\text{bot}}+U_l(\mathcal{T}^{-1}(u(t,x)))}^{h_{\text{soil}}}\partial_t p\,dz$$

$$+\,\text{div}_x\Big(\int_{h_{\text{bot}}+U_l(\mathcal{T}^{-1}(u))}^{h_{\text{soil}}} q\,dz\Big) + q_{|z=h_{\text{soil}}^+}\cdot\vec{\nu}\Big) \quad \text{dans } (0,T)\times\Omega_x, \tag{1.178}$$

$$\nabla u\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times\partial\Omega_x, \qquad u(0,x) = \frac{(h_0(x) - h_{\text{bot}}(x))^2}{2} \quad \text{dans } \Omega_x, \tag{1.179}$$

où P_s est constant par rapport au temps et à l'espace dans ces deux problèmes. La fonction $P_0 \in H^2(\Omega)$ satisfait la condition de compatibilité :

$$P_0(x, h_0) = P_s \quad \text{dans } \Omega_0.$$

De même, nous supposons que $h_0 \in L^\infty(\Omega_x)$ est tel que :

$$h_{\text{bot}} + \delta \leq h_0 \leq h_{\text{soil}} \quad \text{p.p} \quad \text{dans } \Omega_x. \tag{1.180}$$

Le terme source Q est une fonction donnée dans $L^2(0,T;H(\Omega))$. On suppose que les termes sources F et $P_{|\Gamma_{\text{soil}}}$ sont des fonctions données dans l'espace $L^2(0,T,L^2(\Omega_x))$.

Pour les systèmes paraboliques ci-dessus, nous énonçons et prouvons les résultats d'existence suivants :

Théorème 1 *Supposons qu'il existe deux nombres réels θ_- et κ_- tels que :*

$$\theta(x) \geq \theta_- > 0 \quad \forall x \in \mathbb{R} \qquad \text{et} \qquad \kappa(x) \geq \kappa_- > 0 \quad \forall x \in \mathbb{R}^+. \tag{1.181}$$

Alors le système (1.172)-(1.173), (1.174)-(1.175) admet une solution faible (p,u) satisfaisant :

(a) la fonction $p \in L^2(0,T;H^1(\Omega)) \cap L^2(0,T;(H^1(\Omega))')$ est solution de (1.172)-(1.173);

(b) la fonction $u \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ est solution de (1.174)-(1.175). De plus, $u(t,x) \geq 0$ p.p dans $[0,T] \times \Omega_x$.

Également, le système (1.176)- (1.177),(1.178)-(1.179) admet une solution faible (p,u) satisfaisant :

(a) la fonction $p \in L^2(0,T;H^1(\Omega)) \cap L^2(0,T;(H^1(\Omega))')$ est solution de (1.176)- (1.177);

(b) la fonction $u \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ est solution de (1.178)-(1.179). De plus, $u(t,x) \geq 0$ p.p dans $[0,T] \times \Omega_x$.

Corollaire 1 *Supposons qu'il existe deux nombres réels θ_- et κ_- vérifiant (1.181). Alors chaque modèle $(\mathcal{M})$ et $(\mathcal{N})$ admet une solution faible (P,h) telle que :*

(a) la fonction $P \in L^2(0,T;H^1(\Omega)) \cap L^2(0,T;(H^1(\Omega))')$;

(b) la fonction $h \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ et $h(t,x) \in [h_{bot}+\delta, h_{soil}]$ p.p dans $[0,T] \times \Omega_x$.

La preuve du Corollaire 1 est une conséquence directe du Théorème 1 puisque nous revenons au problème original $(\mathcal{M})$ en considérant la transformée inverse $\mathcal{P}^{-1}$ (et la transformée inverse $\mathcal{T}^{-1}$ dans le cas du modèle $(\mathcal{N})$).

La section suivante est consacrée à donner les principaux axes de la preuve du thèorème 1. On renvoie au chapître 3 (pour le modèle $(\mathcal{M})$) et au chapître 4 (pour le modèle $(\mathcal{N})$) pour une démonstration plus détaillée.

1.4.7 Trame de la preuve du Théorème 1

Ainsi que nous l'avons déjà mentionné, les deux problèmes consistent en des systèmes couplés fortement non linéaires, nous appliquons donc une approche à point fixe pour les résoudre en deux étapes. Dans la première étape, nous découplons chaque système et appliquons un théorème de Schauder à point fixe pour établir un résultat d'existence et d'unicité pour chaque équation découplée et régularisée. Ensuite, nous établissons des résultats de compacité qui nous permettent de prouver l'existence globale en temps du probléme initial. L'une des principales difficultés de l'étude est que nous travaillons sur des domaines dépendant du temps. Cette difficulté est résolue en utilisant les travaux de Lions et Mignot pour les équations paraboliques sur des domaines non cylindriques. Cela consiste à prolonger convenablement la solution en dehors du domaine variable, nous ramenant ainsi à un domaine fixe ([57, 58]).

Tout d'abord, on réduit la condition aux limites sur l'interface Γ_t des systèmes (1.172)-(1.173) et (1.176)- (1.177) à une condition aux limites de Dirichlet homogène. Pour ce faire, on définit $\bar{p} = p - \mathcal{P}(P_s)$. Puisque $\mathcal{P}(P_s)$ est une constante, alors :

- le système (1.172)-(1.173) devient :

$$\begin{aligned}
&\partial_t \bar{p} - \nabla\cdot\Big(\bar{\tau}(\bar{p}))\,\nabla\bar{p}\Big) - \frac{\rho_0\, g}{\mu}\nabla\cdot\Big(\bar{\kappa}(\bar{p})K_0\,\vec{e_3}\Big) = Q \quad \text{dans } \mathcal{O}_T,\\
&\bar{p}\Big|_{\Gamma_t} = 0 \quad \text{dans } (0,T), \qquad \nabla\Big(\mathcal{P}^{-1}\Big(\bar{p}+\mathcal{P}(P_s)\Big) + \rho_0\, g\, z\Big)\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times(\Gamma_{\text{soil}}\cup\Gamma_{\text{ver}}),\\
&\bar{p}(0,x,z) = \mathcal{P}(P_0)(x,z) - \mathcal{P}(P_s) \quad \text{dans } \Omega_0,
\end{aligned}$$

- le système (1.176)- (1.177) devient :

$$\begin{aligned}
&\partial_t \bar{p} - \nabla\cdot\Big(\bar{\sigma}(\bar{p}))\,B\,\nabla\bar{p}\Big) - \nabla\cdot\Big(\bar{\kappa}(\bar{p})B\vec{e_3}\Big) = 0 \quad \text{dans } \mathcal{O}_T,\\
&\bar{p}\Big|_{\Gamma_t} = 0 \quad \text{dans } (0,T), \qquad \nabla\Big(\mathcal{P}^{-1}\Big(\bar{p}+\mathcal{P}(P_s)\Big) + \rho g z\Big)\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times\Gamma_{ver},\\
&a\,P + \nabla\Big(\mathcal{P}^{-1}(\bar{p}+\mathcal{P}(P_s)) + b\rho g z\Big)\cdot\vec{\nu} = F \quad \text{sur } (0,T)\times\Gamma_{\text{soil}},\\
&\bar{p}(0,x,z) = \mathcal{P}(P_0)(x,z) - \mathcal{P}(P_s) \quad \text{dans } \Omega_0,
\end{aligned}$$

où

$$\bar{\tau}(\bar{p}) = \tau\Big(\bar{p}+\mathcal{P}(P_s)\Big), \quad \bar{\sigma}(\bar{p}) = \sigma\Big(\bar{p}+\mathcal{P}(P_s)\Big) \quad \text{et} \quad \bar{\kappa}(\bar{p}) = \kappa\circ\mathcal{P}^{-1}\Big(\bar{p}+\mathcal{P}(P_s)\Big).$$

On remarque donc qu'en renommant simplement les fonctions τ, σ et κ, on revient au cas $\mathcal{P}(P_s) = 0$ sur Γ_t dans chaque problème. Donc, dans la suite, nous omettons l'indice " $\bar{\cdot}$ " dans les systèmes précédents et nous considérons les systèmes originaux (1.172)-(1.173) et (1.176)- (1.177) avec $\mathcal{P}(P_s) = 0$. Alors nous obtenons respectivement :

$$\partial_t p - \nabla\cdot\Big(\tau(p)\,\nabla p\Big) - \frac{\rho_0\, g}{\mu}\nabla\cdot\Big(\kappa(\mathcal{P}^{-1}(p))K_0\,\vec{e_3}\Big) = Q \quad \text{dans } \mathcal{O}_T, \tag{1.182}$$

$$p\Big|_{\Gamma_t} = 0 \quad \text{dans } (0,T), \qquad \nabla\Big(\mathcal{P}^{-1}(p) + \rho_0\, g\, z\Big)\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times(\Gamma_{\text{soil}}\cup\Gamma_{\text{ver}}),$$

$$p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{dans } \Omega_0. \tag{1.183}$$

et

$$\partial_t p - \nabla\cdot\Big(\sigma(p)\,B\,\nabla p\Big) - \nabla\cdot\Big(\kappa(\mathcal{P}^{-1}(p))B\vec{e_3}\Big) = 0 \quad \text{dans } \mathcal{O}_T, \tag{1.184}$$

$$p\Big|_{\Gamma_t} = 0 \quad \text{dans } (0,T), \qquad \nabla\Big(\mathcal{P}^{-1}(p) + \rho g z\Big)\cdot\vec{\nu} = 0 \quad \text{sur } (0,T)\times\Gamma_{ver},$$

$$a\,P + \nabla\Big(\mathcal{P}^{-1}(p) + b\rho g z\Big)\cdot\vec{\nu} = F \quad \text{sur } (0,T)\times\Gamma_{\text{soil}}, \qquad p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{dans } \Omega_0. \tag{1.185}$$

Définition 2 *On appelle solution faible du problème (1.182)-(1.183) toute solution $p \in W(0,T,\Omega)$ telle que :*

1. *$p = 0$ dans $\Omega\setminus\Omega_t$, $\forall t\in(0,T)$,*
2. *la solution p satisfait la formulation faible dans $\mathcal{O}_T$: $\forall\phi\in\mathcal{A}(\mathcal{O})$ (nul sur l'interface Γ_t)*

$$\langle\partial_t p,\phi\rangle_{F',F} + \int_0^T\Big(\int_{\Omega_t}\Big(\tau(p)\nabla p + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(p))K_0\,\vec{e_3}\Big)\cdot\nabla\phi\Big) - Q\phi\Big)\,dt = 0, \tag{1.186}$$

$$p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{dans } \Omega_0. \tag{1.187}$$

Définition 3 *On appelle solution faible du problème* (1.178)-(1.179) *toute solution* $u \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ *qui satisfait la formulation faible :* $\forall \phi \in L^2(0,T;H^1(\Omega_x))$

$$S_0\langle \partial_t u, \phi\rangle_{(H^1)',H^1} + \int_0^T \int_{\Omega_x} \Big(\frac{\mathcal{T}^{-1}(u)}{\mathcal{T}'(\mathcal{T}^{-1}(u))} \tilde{J}\nabla' u \cdot \nabla' \phi\Big) dx\, dt$$
$$= -\int_0^T \int_{\Omega_x} \chi(u) \Big(\Big(\int_{h(t,x)}^{h_{\text{soil}}} \partial_t \bar{p}\, dz + aP\big|_{\Gamma_{\text{soil}}} - F \Big)\phi - \Big(\int_{h(t,x)}^{h_{\text{soil}}} q dz \Big) \cdot \nabla' \phi \Big) dx\, dt, \tag{1.188}$$

$$u(0,x) = \frac{(h_0(x) - h_{\text{bot}}(x))^2}{2} \quad \textit{dans } \Omega_x, \tag{1.189}$$

où $h(t,x) := h_{\text{bot}} + U_l(\mathcal{T}^{-1}(u(t,x)))$ *et* $q = -\kappa(P)\, B\nabla\Big(\frac{P}{\rho g} + z\Big)$.

Nous déduisons directement la formulation faible précédente de l'équation (1.178) en gardant :

$$q \cdot \vec{\nu} = 0 \quad \text{sur} \quad (0,T) \times \Gamma_{\text{ver}}.$$

Définition 4 *On appelle solution faible du problème* (1.184)-(1.185) *toute solution* $p \in W(0,T,\Omega)$ *telle que :*

1. $p = 0$ *dans* $\Omega \setminus \Omega_t$, $\forall t \in (0,T)$,
2. *la solution* p *satisfait la formulation faible dans* $\mathcal{O}_T$, $\forall \phi \in \mathcal{A}(\mathcal{O})$ *(nul sur l'interface* Γ_t*)*

$$\langle \partial_t p, \phi\rangle_{F',F} + \int_0^T \Big(\int_{\Omega_t} \Big(\sigma(p) B\nabla p + \kappa(\mathcal{P}^{-1}(p)) B\vec{e_3} \Big) \cdot \nabla \phi \Big) \Big) dt$$
$$= \int_0^T \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}})\, \phi_{|\Gamma_{\text{soil}}}\, dx\, dt, \tag{1.190}$$

$$p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \textit{dans } \Omega_0. \tag{1.191}$$

Le but maintenant est de présenter le cadre pour appliquer le théorème de point fixe de Schauder (voir [63, 64]).
Pour la stratégie de point fixe, on introduit deux sous-ensembles convexes (W_1, W_2) de $W(0,T,\Omega_x) \times W(0,T,\Omega)$ tels que :

$$W_1 := \{u \in W(0,T,\Omega_x);\quad u(0) = u_0,\quad \|u\|_{L^2(0,T;H^1(\Omega_x))} \le C_u \quad \text{et} \quad \|u\|_{L^2(0,T;(H^1(\Omega_x))')} \le C_u'\}$$
$$W_2 := \{p \in W(0,T,\Omega);\quad p(0) = p_0,\quad \|p\|_{L^2(0,T;H^1(\Omega))} \le C_p \quad \text{et} \quad \|p\|_{L^2(0,T;(H^1(\Omega))')} \le C_p'\},$$

où, pour chaque problème, les constantes (C_p, C_p') et (C_u, C_u') sont définies, à partir des preuves (détaillées dans les chapitres 3 et 4) du Lemme 3 et de la Proposition 4 donnés ci-après.
Soit $(\bar{u}, \bar{p}) \in W_1 \times W_2$, on commence par considérer les solutions uniques des problèmes linéarisés suivants :

$$\frac{S_0}{2}\partial_t u - \frac{\tilde{K}}{2}\nabla' \cdot (\nabla' u) = -\chi(u) \int_{\bar{h}(t,x)}^{h_{\text{soil}}(x)} \Big(\frac{\partial \bar{p}}{\partial t} - Q\Big) dz \quad \text{dans } (0,T) \times \Omega_x, \tag{1.192}$$

$$\nabla u \cdot \vec{\nu} = 0 \quad \text{sur } (0,T) \times \partial\Omega_x \qquad \text{et} \qquad u(0,x) = (h_0(x) - h_{\text{bot}}(x))^2 \quad \text{dans } \Omega_x, \tag{1.193}$$

où $\bar{h}(t,x) := T_l(\bar{u}(t,x))$, et

$$S_0 \partial_t u - \nabla' \cdot \Big(\frac{\mathcal{T}^{-1}(\bar{u})}{\mathcal{T}'(\mathcal{T}^{-1}(\bar{u}))} \tilde{J}\nabla' u \Big)$$
$$= -\chi(u) \Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}} \partial_t \bar{p}\, dz + \text{div}_x \Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}} \bar{q}\, dz \Big) + aP\big|_{\Gamma_{\text{soil}}} - F \Big) \quad \text{dans } (0,T) \times \Omega_x, \tag{1.194}$$

$$\nabla u \cdot \vec{\nu} = 0 \quad \text{sur } (0,T) \times \partial\Omega_x, \qquad u(0,x) = \frac{(h_0(x) - h_{\text{bot}}(x))^2}{2} \quad \text{dans } \Omega_x, \tag{1.195}$$

où $\bar{h}(t,x) := h_{\text{bot}} + U_l(\mathcal{T}^{-1}(\bar{u}(t,x)))$.

Remarque 5 1. *Pour les deux problèmes (1.192)-(1.193) et (1.194)-(1.195), il faut préciser le sens du terme* $\int_{\bar{h}(t,x)}^{h_{soil}(x)} \frac{\partial \bar{p}}{\partial t}\, dz$ *:*

$$\int_{\bar{h}(t,x)}^{h_{soil}(x)} \frac{\partial \bar{p}}{\partial t}\, dz = \int_{h_{bot}(x)}^{h_{soil}(x)} \chi_{z\geq \bar{h}(t,x)} \frac{\partial \bar{p}}{\partial t}\, dz$$

est la fonction de $(H^1(\Omega_x))'$ *telle que :* $\forall v \in H^1(\Omega_x) \subset H^1(\Omega)$,

$$\langle \int_{\bar{h}(t,x)}^{h_{soil}} \frac{\partial \bar{p}}{\partial t}\, dz,\, v\rangle_{(H^1(\Omega_x))',H^1(\Omega_x)} = \langle \frac{\partial \bar{p}}{\partial t}, \chi_{z\geq \bar{h}(t,x)}\, v\rangle_{(H^1(\Omega))',H^1(\Omega)}.$$

2. *Grâce au changement de variable* $u = (h - h_{bot})^2$*, l'équation (1.110) (qui est non linéaire et dégénéré en espace et en temps) possède maintenant une structure parabolique.*
3. *Dans le système (1.194)-(1.195), compte tenu de la condition aux limites de Robin sur la frontière* $z = h_{\text{soil}}$*, nous remplaçons le flux* $q_{|z=h_{\text{soil}}^+} \cdot \vec{\nu}$ *par* $F - aP_{|\Gamma_{\text{soil}}}$*. Ces données expriment le terme source et les échanges entre l'eau de surface et l'aquifère. Notons également que puisque* h_{soil} *est supposé constant, le vecteur normal unitaire* $\vec{\nu}$ *correspond donc à* $\vec{e_3}$*. Dans ce qui suit, on note par* $F_R = F - aP_{|\Gamma_{\text{soil}}} \in L^2(0,T,L^2(\Omega_x))$ *le résultat de ces deux entrées externes.*

Lemme 3 *Soit* $h_0 \in L^\infty(\Omega_x)$ *satisfaisant (1.180). Alors pour chaque problème (1.192)-(1.193) et (1.194)-(1.195), il existe une unique solution faible* $u \in W(0,T,\Omega_x)$ *telle que :*

$$\|u\|_{L^2(0,T;H^1(\Omega_x))} \leq C_u \quad \textit{et} \quad \|u\|_{L^2(0,T;(H^1(\Omega_x))')} \leq C_u',$$

où C_u *et* C_u' *ne dépendent que des données du problème.*
De plus, $u \geq 0$ *p.p dans* $[0,T]\times\Omega_x$*.*
On note u_1 *la solution de (1.192)-(1.193) et* u_2 *la solution de (1.194)-(1.195).*

Grâce aux changements de variables, nous nous sommes ramenés dans les deux cas, à des équations paraboliques auxquelles nous pouvons appliquer les théories générales pour l'existence et l'unicité de la solution (cf. [67, 68]). Les estimations en norme $L^2(0,T;H^1(\Omega_x))$ et $L^2(0,T;(H^1(\Omega_x))')$ s'obtiennent de manière classique par des estimations d'énergie.

Les résultats énoncés dans le Lemme 1 nécessitent d'avoir des domaines non cylindriques réguliers notamment avec des frontières suffisamment régulières (de classe C^1 par morceaux comme mentionné par Mignot). Comme, dans notre problème, on ne peut garantir autant de régularité à l'interface h (qui est dans $W(0,T,\Omega_x)$, un processus de régularisation est utilisé pour placer notre étude dans le cadre de Mignot [58].
On régularise donc h par une convolution en espace. Soit $\psi \in C^\infty(\mathbb{R}^2)$, $\psi \geq 0$, avec support dans la boule unité tel que $\int_{\mathbb{R}^2} \psi(x)\, dx = 1$. Pour $\eta > 0$ assez petit, on pose $\psi_\eta(x) = \psi(x/\eta)/\eta^2$. On prolonge h par zéro à l'extérieur de Ω_x, donc on a $h \in C([0,T];L^2(\mathbb{R}^2)) \cap W(0,T,\mathbb{R}^2)$. On définit ainsi $\tilde{h}$ par le produit de convolution par rapport à la variable d'espace :

$$\tilde{h} = \psi_\eta * h.$$

Sa restriction à Ω_x est notée de la même manière. De ce fait, $\tilde{h} \in C^\infty(\bar{\Omega}_x)$, et quand $\eta \to 0$, on a :

$$\tilde{h} \to h \quad \text{fortement dans} \quad C([0,T];L^2(\Omega_x)) \cap L^2(0,T,H^1(\Omega_x)).$$

Nous remplaçons h par $\tilde{h}$ dans les équations (1.182)-(1.183) et (1.184)-(1.185), (la substitution apparaît dans le domaine d'intégration spatiale Ω_t).

Soient $\bar{p} \in W_1$ et $\tilde{h}\,(=\psi_\eta * h) \in C^\infty(\bar{\Omega}_x)$ où h est donné par le Lemme 3. Nous considérons donc les problèmes suivants, linéarisés et régularisés dans Ω_T :

- Trouver $p_\eta \in W(0,T,\Omega)$ tel que $\forall \phi \in \mathcal{A}(\mathcal{O})$ (nul sur l'interface Γ_t définie par $\tilde{h}$)

$$\langle \partial_t p_\eta, \phi\rangle_{F',F} + \int_0^T \left(\int_{\Omega_t} \left(\tau(\bar{p}) \nabla p_\eta + \frac{\rho_0\, g}{\mu} \kappa(\mathcal{P}^{-1}(\bar{p})) K_0\, \vec{e_3} \right) \cdot \nabla \phi \right) - Q\phi \Big) \, dt = 0, \tag{1.196}$$

$$p_\eta = 0 \quad \text{dans } \Omega \setminus \Omega_t,\ \forall t \in [0,T] \quad \text{et} \quad p_\eta(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{dans } \Omega_0. \tag{1.197}$$

- Trouver $p_\eta \in W(0,T,\Omega)$ tel que $\forall \phi \in \mathcal{A}(\mathcal{O})$ (nul sur l'interface Γ_t définie par $\tilde{h}$)

$$\begin{aligned} \langle \partial_t p_\eta, \phi\rangle_{F',F} + \int_0^T \left(\int_{\Omega_t} \left(\sigma(\bar{p})\, B \nabla p_\eta + \kappa(\mathcal{P}^{-1}(\bar{p})) B\, \vec{e_3} \right) \cdot \nabla \phi \right) \Big) \, dt = 0, \\ = \int_0^T \int_{\Omega_x} (F - a P_{|_{\Gamma_{\text{soil}}}})\, \phi dx\, dt, \end{aligned} \tag{1.198}$$

$$p_\eta = 0 \quad \text{dans } \Omega \setminus \Omega_t,\ \forall t \in [0,T] \qquad p_\eta(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{dans } \Omega_0. \tag{1.199}$$

Nous admettons la Proposition 4 dont la preuve est donnée à la fin du chapitre 3 (pour le modèle $(\mathcal{M})$) et à la fin du chapitre 4 avant la partie Annexe (pour le modèle $(\mathcal{N})$).

Proposition 4 *Pour tout $\eta > 0$, il existe une fonction unique p_η dans $W(0,T,\Omega)$ solution de (1.196)-(1.197). Elle vérifie les estimations uniformes :*

$$\|p_\eta\|_{L^2(0,T;H^1(\Omega))} \le C_p \quad \textit{et} \quad \|p_\eta\|_{L^2(0,T;(H^1(\Omega))')} \le C_p', \tag{1.200}$$

où C_p et C_p' ne dépendent que des données du problème d'origine (1.172)-(1.173).
De même pour le système (1.198)-(1.199).

La preuve de cette proposition fait appel à la théorie des équations paraboliques définies sur des ouverts non cylindriques qui a été introduite par Lions et Mignot. L'idée principale est de prolonger par zéro la solution p_η en dehors de la frontière libre. Il est donc essentiel d'avoir une condition de Dirichlet homogème sur cette frontière. On est ensuite ramené à l'étude d'une équation parabolique sur un domaine fixe pour une fonction p_η définie sur tout le domaine $(0,T)\times\Omega$. On définit une famille de problèmes pénalisés qui sont des problèmes paraboliques linéaires dans le domaine cylindre $(0,T)\times\Omega$, et dont la solution restreinte à l'ensemble $\mathcal{O}_T$ convergera vers la solution p de l'équation linéarisée déduite de (1.172) (respectivement de (1.198)).

Il s'agit maintenant de prouver l'existence d'une solution pour le problème régularisé. Dans la suite, nous omettons l'indice η dans p_η (et aussi dans u_η).

Soient $(\bar{u}_1,\bar{p}_1) \in W(0,T,\Omega_x)\times W(0,T,\Omega)$ la solution unique de (1.192)-(1.193) et (1.196)-(1.197) et $(\bar{u}_2,\bar{p}_2) \in W(0,T,\Omega_x)\times W(0,T,\Omega)$ la solution unique de (1.194)-(1.195) et (1.198)-(1.199). Le Lemme 3 et la Proposition 4 permettent de définir deux applications $\mathcal{F}_1$ et $\mathcal{F}_2$ telles que :

$$\begin{array}{rccc} \mathcal{F}_1 : & W(0,T,\Omega_x)\times W(0,T,\Omega) & \longrightarrow & W(0,T,\Omega_x)\times W(0,T,\Omega) \\ & \bar{u}_1,\bar{p}_1 & \longmapsto & \mathcal{F}_1(\bar{u}_1,\bar{p}_1) \end{array} \quad \Big| \quad \mathcal{F}_1(\bar{u}_1,\bar{p}_1) = (u_1,p_1) \tag{1.201}$$

et

$$\begin{array}{rccc} \mathcal{F}_2 : & W(0,T,\Omega_x)\times W(0,T,\Omega) & \longrightarrow & W(0,T,\Omega_x)\times W(0,T,\Omega) \\ & \bar{u}_2,\bar{p}_2 & \longmapsto & \mathcal{F}_2(\bar{u}_2,\bar{p}_2) \end{array} \quad \Big| \quad \mathcal{F}_2(\bar{u}_2,\bar{p}_2) = (u_2,p_2) \tag{1.202}$$

La fin de la présente sous-section est consacrée à la preuve de l'existence d'un point fixe pour $(\mathcal{F}_1,\mathcal{F}_2)$ dans un sous ensemble approprié. Ce résultat découlera directement de l'application du Théorème du

point fixe de Schauder (see [67, 68]). Le Lemme 3 et la Proposition 4 nous garantissent l'existence de deux convexes fermés bornés non vides de $W(0,T,\Omega_x)\times W(0,T,\Omega)$ et nous permettent aussi de montrer la continuité séquentielle au sens faible des applications $(\mathcal{F}_1,\mathcal{F}_2)$. On peut ainsi énoncer le résultat d'existence suivant :

Lemme 4 *Soient $(\bar{u}_1,\bar{p}_1)\in W(0,T,\Omega_x)\times W(0,T,\Omega)$ la solution unique de (1.192)-(1.193) et (1.196)-(1.197) et $(\bar{u}_2,\bar{p}_2)\in W(0,T,\Omega_x)\times W(0,T,\Omega)$ la solution unique de (1.194)-(1.195) et (1.198)-(1.199). Alors :*

- *Il existe deux ensembles $\mathcal{C}_1$ et $\mathcal{C}_2$ dans $W(0,T,\Omega_x)\times W(0,T,\Omega)$ non vides, fermés, convexes et bornés satisfaisant $\mathcal{F}_1(\mathcal{C}_1)=\mathcal{C}_1$ et $\mathcal{F}_2(\mathcal{C}_2)=\mathcal{C}_2$;*
- *Les application $\mathcal{F}_1$ et $\mathcal{F}_2$ définie par (1.201) et (1.202) respectivement sont faiblement séquentiellement continues dans $W(0,T,\Omega_x)\times W(0,T,\Omega)$.*
- *Il existe $(u_1,p_1),(u_2,p_2)\in W_1\times W_2$ tels que $\mathcal{F}_1(u_1,p_1)=(u_1,p_1)$ et $\mathcal{F}_2(u_2,p_2)=(u_2,p_2)$.*

On a ainsi obtenu l'existence d'une solution pour le problème régularisé. En passant à la limite lorsque $\eta\to 0$, on obtient l'existence d'une solution pour notre système d'origine, ce qui conclut la démonstration du Théorème 1 .

Chapitre 2

Chimie à l'équilibre thermodynamique

CONVERGENCE ACCELERATION OF ITERATIVE SEQUENCES FOR EQUILIBRIUM CHEMISTRY COMPUTATIONS.

SAFAA AL NAZER, MUSTAPHA JAZAR, AND CAROLE ROSIER

Abstract. The modeling of thermodynamic equilibria leads to complex nonlinear chemical systems which are often solved with the Newton-Raphson method. But this resolution can lead to a non convergence or an excessive number of iterations due to the very ill-conditioned nature of the problem. In this work, we combine a particular formulation of the equilibrium system called the Positive Continuous Fraction method with two iterative methods, Anderson Acceleration method and Vector extrapolation methods (namely the reduced rank extrapolation and the minimal polynomial extrapolation). The main advantage of this approach is to avoid forming the Jacobian matrix. In addition, a strategy is used to improve the robustness of the Anderson acceleration method which consists in reducing the condition number of matrix of the least squares problem in the implementation of the Anderson acceleration so that the numerical stability can be guaranteed. We compare our numerical results with those obtained with the Newton-Raphson method on the Acid Gallic test and the $1D$ MoMas benchmark test case and we show the high efficiency of our approach.

Keywords: Nonlinear systems; Thermodynamic chemistry; Anderson acceleration; Polynomial vector extrapolation.

1. Introduction

In the last decades, reactive transport was considered a major topic in many different fields of science such as combustion, catalysis, fluid mechanics, chemical engineering and geochemistry. Single phase multicomponent reactive flows are modeled by a masse balance law, Darcy's law and equations of state. In the case of equilibrium reactions, mass action laws consist in algebraic equations linking the activities of involved species. The problem of reactive transport is thus modeled by partial differential equations describing the flow coupled with algebraic equations describing chemical reactions. Due to the complexity of systems and the nonlinearity of chemical processes, reactive multicomponent transport results in an important computational requirement. In this context, two numerical strategies are usually used to solve this system : the global implicit algorithm (GIA) and the sequential iterative (and non-iterative) algorithm (SIA), also called operator splitting approach (see for instance references [2, 29, 38, 45]). The global implicit algorithm solves at each time step the complete nonlinear system resulting from the direct substitution of the chemical equations in the transport equations while the operator splitting approach solves sequentially transport equations and biogeochemical reactions. Results of recent comparisons between GIA and SIA obtained by different teams are in good agreement ([14]), such as those given in ([1]) for which a fully implicit finite volume method has been developed and implemented in the framework of the parallel open-source platform Dumu^X ([19, 21]). These different benchmarks have shown that the precision of sequential approaches is comparable with that of global approaches and that global approaches are now more efficient than originally believed (even if in some work, it is mentioned that the global approach is much more expensive in terms of computation time and storage than the operator splitting approach cf [52]). Each of these methods has qualities and drawbacks but regardless of the approach, a nonlinear problem must be solved by a fixed point method and the Newton Raphson method is often used for this numerical resolution. However, the resolution of such nonlinear systems, especially due to chemical processes can yield a non convergence or an excessive number of iterations due to the very ill-conditioned nature of the problem. The goal of this work is to suggest new powerful algorithms (in terms of CPU time et stability) which will allow to deal with these stiff problems.

In thermodynamic terms, a chemical equilibrium calculation, which aims to find the minimum value for the Gibbs free energy, can be carried out through one of the following ways: by minimizing a free energy function or by solving a set of nonlinear equations consisting of equilibrium constants and mass balance constraints. Note that, in the petroleum industry context, recent alternative approach ([54]) studies phase equilibrium under a fixed volume rather than fixed pressure and minimizes Helmholtz free energy instead of Gibbs free energy. Finally, recent works use efficient deep learning algorithms to estimate the thermodynamic equilibrium

Date: 24 - Juin - 2020.

states of realistic reservoir fluids with a large number of components thus allowing to accelerate phase equilibrium calculations. More precisely, a simple acceleration strategy reduces the number of components in the fluid mixture improving the efficiency of algorithms without compromising the accuracy of equations of states (see [55, 56]). These methods are thermodynamically equivalent, but the main disadvantage of using a free energy database is that these values are not nearly as reliable as directly measured equilibrium constants. As the accuracy of results of chemical solvers is particularly required, especially if one wants to integrate them in SIA methods, we are going to focus on the numerical resolution of nonlinear equations describing thermodynamic equilibria.

Many mathematical methods were tested to solve the set of nonlinear algebraic equations describing thermodynamic equilibrium: Zero-order methods such as the continuous fractions method [51], the Simplex method ([32]) which do not use the derivative of the objective function. The latter methods converge more slowly [31], 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 ([6, 33]). As mentioned above, the Newton-Raphson method is the most used to compute thermodynamic equilibrium or more generally to solve the set of nonlinear equations. For example, let us quote software such as HYDROGEOCHEM [53], DUNE [21], IMPACT [26], CHESS [47], or PHREEQC [33], with the difficulty that the Jacobian matrix has to be computed, stored, factored and is usually very ill-conditioned, which requires preconditioning procedures [15, 16]. This can become problematic for large problems. In addition, the localization of the initial data in any algorithm of Newton type is a recurrent difficulty which slows down and even prevents the convergence of the algorithm. Finally, even small or very small chemical systems (4×4 to 20×20, occasionally larger) can be very ill-conditioned (condition number up to 10^{100}) as it is shown in [28].

To overcome this problematic, it is more effective to solve the chemical equilibrium problem through other iterative methods not requiring the calculation of the Jacobian matrix, by first transforming it into an appropriated fixed point problem. We especially focused on three iterative acceleration methods: the *Anderson Acceleration* method (AA) originating in [3] and *Vector-Extrapolation methods*, mainly the two polynomial-type methods, which include the *Reduced-Rank Extrapolation* (RRE) of Eddy [17] and MeSina [30], and *Minimal-Polynomial Extrapolation* (MPE) of Cabay and Jackson [9]. To our knowledge, these methods have never been applied to the resolution of thermodynamic equilibria. Moreover, their efficiency is improved by combining them with a particular formulation of the equilibrium system: the positive continuous fractions method PCF. Usually, continuous fractions method is used for preconditioning the Newton-Raphson method for major species (as in the PHREEQC [33]) or to reduce the difficulties due to the lack of global convergence of Newton's method, if the initial condition is not sufficiently close to the solution (see [12]). The direct combination of PCF method with AA, RRE or MPE presented in this work provides very efficient and robust algorithms with a super linear or quadratic convergence from any arbitrary initial data. Let us now briefly describe these three iterative methods which are part of a general framework of Shanks sequence transformations (cf.[7]). Anderson acceleration is related to multisecant methods (extensions of quasi-Newton methods involving multiple secant conditions); actually, Eyert [18] proves that it is equivalent to the so-called "bad" Broyden's method [8], and a similar analysis is done by Fang and Saad [22] and Rohwedder and Schneider [36]. As for linear systems, if $m_k = k$ for each k then Anderson acceleration is essentially equivalent to the generalized minimal residual (GMRES) method [37], as shown by Potra and Engler [34], Rohwedder and Schneider [36], and Walker and Ni [49]. For nonlinear problems Rohwedder and Schneider [36] show that Anderson acceleration is locally linearly convergent under certain conditions. In addition to the previous convergence analysis results, the recent work by Toth and Kelley [46] concerning Anderson acceleration with $m_k = \min(m, k)$, for a fixed m, applied to contractive mappings should be mentioned. Regarding Vector-Extrapolation methods, the aim of such methods is to transform a sequence of vectors generated by some process to a new one with the goal to converge faster than the initial sequence towards the sought limit solution. An example to these vector sequences is those which are obtained from iterative solution of linear and nonlinear systems of equations. These methods can be classified into two main categories: the polynomial methods and the ϵ-algorithms. There exists many polynomial extrapolation methods but, in this paper, we will be interested in the minimal polynomial extrapolation method (MPE) of Cabay and Jackson [9] as well as the reduced rank extrapolation method (RRE) of Eddy [17] and Mesina [30]. These methods do not require any explicit knowledge of how the sequence is generated, and consequently can be directly applied for solving linear and nonlinear systems. They are especially effective in the nonlinear case.

In the first part of this work, a brief description of the chemical context and the chemical modeling strategy are given. Based on the Positive Continuous Fractions method [12], it is transformed into a fixed point problem. In the second part, a survey of Anderson Acceleration method and of the two most efficient and widely used vector extrapolation methods MPE and RRE is given. By derivating these methods, stable and efficient algorithms are obtained. In the third part, numerical results for solving thermodynamic equilibrium problem by Anderson Acceleration, MPE and RRE methods are detailed. Third part contains a brief description of two chemical tests: Gallic acid test and MoMas easy test case. Using the data of this tests presented by their Morel's Tables, three iterative methods mentioned for solving the fixed point problem of chemical equilibrium are applied and numerical results for each test and each method are given. In the fourth part, a comparison between these results and other results is presented in order to prove the effectiveness of methods used in this work to solve the problem of chemical equilibrium in porous media.

2. Description and modeling of chemistry

In this section, chemical model studied in this work is described. Consider a set of n_e chemical species $(\mathcal{E}_j)$, $j = 1, ..., n_e$ linked by n_r reactions such that $n^r \leq n^e$

$$\sum_{j=1}^{n_e} \tilde{\mu}_{ij} \mathcal{E}_j \rightleftarrows 0, \; i = 1, ..., n_r, \tag{2.1}$$

where $\tilde{\mu}_{ij}$ is the stoichiometric matrix of species $\mathcal{E}_j$ in the reaction i. (2.1) can be written in matrix form

$$\tilde{\mu}\mathcal{E} \rightleftarrows 0.$$

After substitution and relabeling, each reaction can be written in a form giving rise a single distinct product per reaction. It is natural to assume that the stoichiometric matrix $\tilde{\mu}$ is of full rank n_r. So $\tilde{\mu} = [-I_{n_r} \quad \mu]$ can be written in the echelon form, where I_{n_r} is the identity matrix of size n_r. The chemical system is then written (after a possible numbering) in the form

$$\mathcal{C}_i \rightleftarrows \sum_{j=1}^{n_e - n_r} \mu_{ij} \mathcal{X}_j \quad i = 1, ..., n_r, \tag{2.2}$$

or in matrix form

$$\mathcal{C} \rightleftarrows \mu^T \mathcal{X}$$

where $\mathcal{C}$ (respectively $\mathcal{X}$) are called secondary species (respectively component species). Thus, the equation (2.2) show that the formation of secondary species $\mathcal{C}$ is done from the component species $\mathcal{X}$, in a unique way. The advantage of this approach is that it reduces the size of the chemical system to be solved. *Mobile* and *fixed* species are also distinguished. A species is said to be mobile (m) if it belongs to a mobile phase, fixed (f) if it belongs to the fixed phase and precipitated if it is mineral (π). Using the following notations:

- X: subset of mobile component species of cardinal n_{pm},
- S: subset of fixed component species of cardinal n_{pf},
- C: subset of mobile secondary species of cardinal n_{sm},
- CS: subset of fixed secondary species of cardinal n_{sf},
- π: subset of precipitated species of cardinal n_π,
- $\mu^{(C,X)} \in \mathbb{R}^{n_{sm} \times n_{pm}}$: block of the stoichiometric matrix between C and X,
- $\mu^{(\pi,X)} \in \mathbb{R}^{n_\pi \times n_{pm}}$: block of the stoichiometric matrix between π and X,
- $\mu^{(CS,X)} \in \mathbb{R}^{n_{sf} \times n_{pm}}$: block of the stoichiometric matrix between CS and X,
- $\mu^{(CS,S)} \in \mathbb{R}^{n_{sf} \times n_{pf}}$: block of the stoichiometric matrix between CS and S,

Chemical system can be synthesized as follows

$$\begin{pmatrix} & \mu^{(C,X)} & 0 \\ -I_{n_r} & \mu^{(CS,X)} & \mu^{(CS,S)} \\ & \mu^{(\pi,X)} & 0 \end{pmatrix} \begin{pmatrix} C \\ CS \\ \pi \\ X \\ S \end{pmatrix} \rightleftarrows 0,$$

with $\mu = \begin{pmatrix} \mu^{(C,X)} & 0 \\ \mu^{(CS,X)} & \mu^{(CS,S)} \\ \mu^{(\pi,X)} & 0 \end{pmatrix}$, $\mathcal{C} = \begin{pmatrix} C \\ CS \\ \pi \end{pmatrix}$, $\mathcal{X} = \begin{pmatrix} X \\ S \end{pmatrix}$ and $\mathcal{E} = \begin{pmatrix} \mathcal{X} \\ \mathcal{C} \end{pmatrix}$

Note that the fixed component species do not take part in the homogeneous reactions which only involve the mobile species and that the precipitation reactions do not involve the fixed species. In this work, chemical systems without precipitated species are considered, i.e ($\pi = \phi$ and $\mu^{(\pi,X)} = 0$). A classic algorithm [12] to describe mineral precipitation or dissolution makes an a priori hypothesis about the existence or non-existence of minerals. In this work, this hypothesis is assumed.

Chemical Reactions In the following, $X = (X_1, ..., X_{n_{pm}})^T$ denote components where $(X_j)_{j=1}^{n_{pm}}$ are the mobile components species and $S = (S_1, ..., S_{n_{pf}})^T$ where $(S_j)_{j=1}^{n_{pf}}$ are the fixed components species. In the same way, $C = (C_1, ..., C_{n_{sm}})^T$ where $(C_i)_{i=1}^{n_{sm}}$ are the mobile secondary species and $CS = (CS_1, ..., CS_{n_{sf}})^T$ where $(CS_i)_{i=1}^{n_{sf}}$ are the fixed secondary species.
Let $\mu_i, i = 1, 3$ be scalars $\mu_1 = \mu^{(C,X)}$, $\mu_2 = \mu^{(CS,X)}$ and $\mu_3 = \mu^{(CS,S)}$. Using these notations, it becomes easy to distinguish chemical reactions as follows:

i: Reactions among mobile species: $\sum_{j=1}^{n_{pm}} \mu_{1,i,j} X_j \rightleftarrows C_i \quad i = 1, ..., n_{sm}$;
ii: Reactions between mobile and fixed species: $\sum_{j=1}^{n_{pm}} \mu_{2,i,j} X_j + \sum_{j=1}^{n_{pf}} \mu_{3,i,j} S_j \rightleftarrows CS_i \quad i = 1, ..., n_{sf}$.

Mass action law The law of mass action describes how to obtain the concentrations of secondary species, given the concentrations of the component species. This law is only valid for a certain type of reaction, including homogeneous reactions. It is assumed during this work that this law is still valid in the case of surface reactions. Since no precipitation phenomena are considered, for each mobile secondary species C_i, the mass action law is

$$\{C_i\} = K_i^m \prod_{k=1}^{n_{pm}} \{X_k\}^{\mu_{1,i,k}}. \tag{2.3}$$

For each fixed secondary species CS_i, the mass action law is

$$\{CS_i\} = K_i^s \prod_{k=1}^{n_{pm}} \{X_k\}^{\mu_{2,i,k}} \prod_{k=1}^{n_{pf}} \{S_k\}^{\mu_{3,i,k}}. \tag{2.4}$$

where $\{C_i\}$ and $\{CS_i\}$ are the activities of each mobile and fixed secondary species given by the mass action law through the activities of each mobile and fixed component species $\{X_k\}$ and $\{S_k\}$. K^m is the equilibrium constant for reactions among mobile species and K^s is the equilibrium constant for sorption reactions.

The relationship between the activity of a species $\mathcal{E}_j$ and its concentration is given by activity coefficient (γ_j) calculated using specific models (Davies, Debye-Huckel, etc.): $\{\mathcal{E}_j\} = \gamma_j [\mathcal{E}_j]$. A solution is said to be *ideal* when the species does not undergo any interaction. In this case, the activity coefficient (γ) is equal to one. This amounts to confusing activity and concentration. During this work, only the case of ideal solutions will be considered, so $[\mathcal{E}_j] = \mathcal{E}_j$.

Mass conservation law Assuming a closed system (without exchange of matter with outside) and all the reactions at equilibrium, then the total quantity of the species $\mathcal{X}_j$ in the system is invariant. This is expressed in terms of the total concentration T_j^m for an aqueous species and the total concentration T_j^s for a sorbed species. The law of conservation (also called Lavoisier's law) can be expressed by the following two relations:

$$\begin{aligned} T_j^m &= X_j + \sum_{i=1}^{n_{sm}} \mu_{1,i,j} C_i + \sum_{i=1}^{n_{sf}} \mu_{2,i,j} CS_i \quad j = 1, ..., n_{pm} \\ T_j^s &= S_j + \sum_{i=1}^{n_{sm}} \mu_{3,i,j} CS_i \quad j = 1, ..., n_{pf}. \end{aligned} \tag{2.5}$$

The two relationships in (2.5) introduce a distinction between the concentration of the component species and the concentrations of the other secondary species. This distinction is not necessarily necessary. Reactions (2.6) for each component species can be quite considered, which has a stoichiometric coefficient equal to one and an equilibrium constant equal to one,

$$X_j \rightleftarrows C_i \quad \text{and} \quad S_j \rightleftarrows CS_i. \tag{2.6}$$

Then (2.5) is written more simply

$$\begin{aligned} T_j^m &= \sum_{i=1}^{n_{sm}} \mu_{1,i,j} C_i + \sum_{i=1}^{n_{sf}} \mu_{2,i,j} CS_i \quad j = 1, ..., n_{pm} \\ T_j^s &= \sum_{i=1}^{n_{sm}} \mu_{3,i,j} CS_i \quad j = 1, ..., n_{pf}. \end{aligned} \tag{2.7}$$

or in matrix form

$$\begin{aligned} T^m &= \mu_1^T.C + \mu_2^T.CS \\ T^s &= \mu_3^T.CS. \end{aligned}$$

2.1. Resolution of the chemical equilibrium

Chemical system By substituting the mass action laws (2.3) and (2.4) into the mass conservation equations (2.7), one can write the equilibrium chemistry like a nonlinear system formed by conservation laws and mass action laws

$$\begin{aligned} T_j^m &= \sum_{i=1}^{n_{sm}} \mu_{1,i,j} \left(K_i^m \prod_{k=1}^{n_{pm}} X_k^{\mu_{1,i,k}} \right) + \sum_{i=1}^{n_{sf}} \mu_{2,i,j} \left(K_i^s \prod_{k=1}^{n_{pm}} X_k^{\mu_{2,i,k}} \prod_{k=1}^{n_{pf}} S_k^{\mu_{3,i,k}} \right) \quad j = 1, ..., n_{pm} \\ T_j^s &= \sum_{i=1}^{n_{sm}} \mu_{3,i,j} \left(K_i^s \prod_{k=1}^{n_{pm}} X_k^{\mu_{2,i,k}} \prod_{k=1}^{n_{pf}} S_k^{\mu_{3,i,k}} \right) \quad j = 1, ..., n_{pf}. \end{aligned} \tag{2.8}$$

This is a system of $(n_{pm} + n_{pf})$ nonlinear algebraic equations with $(n_{pm} + n_{pf})$ unknowns. It is of course not possible (in general) to calculate the exact solution of this system which will be calculated numerically by iterative methods.

A first difficulty in solving (2.8) comes from the fact that the unknowns are concentrations of the component species. These concentrations are likely to vary on several orders of magnitude, and must remain positive to keep their physical significance. These two constraints make numerical resolution difficult. Fortunately, a simple change of variables eliminates these two difficulties, and has been adopted by most computer codes: the logarithms of the concentrations are taken as unknowns. Thus, the concentrations will be automatically positive, and the unknowns of the nonlinear system will keep a reasonable order of magnitude. In this work and computer code, the logarithms at base 10, "$\log_{10}$", of the component concentrations is used as a variable change.

$$\xi_j = \log_{10}(X_j) \quad \text{and} \quad \eta_j = \log_{10}(S_j).$$

We denote by $\mathbf{K}^m = \log_{10}(K^m)$ and $\mathbf{K}^s = \log_{10}(K^s)$.
The consequences of this transformation on the system are limited. The mass action law in the equations (2.3) and (2.4) becomes reformulated, respectively, as

$$C_i = 10^{(\mathbf{K}_i^m + \sum_{k=1}^{n_{pm}} \mu_{1,i,k} \xi_k)} \quad \text{and} \quad CS_i = 10^{(\mathbf{K}_i^s + \sum_{k=1}^{n_{pm}} \mu_{2,i,k} \xi_k + \sum_{k=1}^{n_{pf}} \mu_{3,i,k} \eta_k)}. \tag{2.9}$$

Then, the nonlinear system (2.8) takes the following form

$$\begin{aligned} T_j^m &= \sum_{i=1}^{n_{sm}} \mu_{1,i,j}.10^{(\mathbf{K}_i^m + \sum_{k=1}^{n_{pm}} \mu_{1,i,k} \xi_k)} + \sum_{i=1}^{n_{sf}} \mu_{2,i,j}.10^{(\mathbf{K}_i^s + \sum_{k=1}^{n_{pm}} \mu_{2,i,k} \xi_k + \sum_{k=1}^{n_{pf}} \mu_{3,i,k} \eta_k)} \quad j = 1, ..., n_{pm} \\ T_j^s &= \sum_{i=1}^{n_{sm}} \mu_{3,i,j}.10^{(\mathbf{K}_i^s + \sum_{k=1}^{n_{pm}} \mu_{2,i,k} \xi_k + \sum_{k=1}^{n_{pf}} \mu_{3,i,k} \eta_k)} \quad j = 1, ..., n_{pf}. \end{aligned}$$

or the matrix form

$$\begin{aligned} T^m &= \mu_1^T \times 10^{(\mathbf{K}^m + \mu_1 \times \xi)} + \mu_2^T \times 10^{(\mathbf{K}^s + \mu_2 \times \xi + \mu_3 \times \eta)} \\ T^s &= \mu_3^T \times 10^{(\mathbf{K}^s + \mu_2 \times \xi + \mu_3 \times \eta)}. \end{aligned} \tag{2.10}$$

where the symbol of matrix product is denoted by $\times$. The matrix form (2.10) still can be written in a more reduced manner

$$\mathbf{T} = \mu^T \times 10^{(\mathbf{K} + \mu \times \omega)}, \tag{2.11}$$

where $\mathbf{T} = \begin{pmatrix} T^m \\ T^s \end{pmatrix}$, $\mathbf{K} = \begin{pmatrix} \mathbf{K}^m \\ \mathbf{K}^s \end{pmatrix}$, $\omega = \begin{pmatrix} \xi \\ \eta \end{pmatrix}$ and $\mu = \begin{pmatrix} \mu_1 & 0 \\ \mu_2 & \mu_3 \end{pmatrix}$.
The nonlinear system (2.10) (or (2.11)) corresponds to the *chemical problem* to be solved for ξ and η given T^m and

T^s. The concentrations of the secondary species can then be computed from (2.9). In the sequel, we assumed that this problem always has a unique positive solution (ξ^*, η^*) for all feasible values of the data T^m and T^s. This assumption is true due to the fact that the chemical equilibrium problem is the consequence of the Gibbs free energy minimization problem. The existence follows from the convexity of the energy functional [40]. In addition, a condition is given for the uniqueness of the solution, which is particularly verified in the case of a single-phase system, which covers the cases treated here.

In this work, three different iterative numerical methods are applied to solve the chemical equilibrium problem which are the Anderson Acceleration method and the two polynomial vector extrapolation methods (MPE and RRE). According to their definitions, these methods are used to solve a general fixed point problem of the form $G(Y) = (Y)$ where $G : \mathbb{R}^n \to \mathbb{R}^n$. So it is necessary to write the chemical problem in the form of a fixed point problem.
The numerical method most used in a large number of geochemical codes for the resolution of this nonlinear system is the Newton's method. The thesis [10] by J. Carrayrou contains a comparison of the different methods to solve the problem of chemical equilibrium. His recommendation is to use a combination of Newton's method with a fixed-point method on a particular formulation of the equilibrium system (the PCF positive continuous fraction method). This combination makes it possible to reduce the difficulties due to the lack of total convergence of the method of Newton, if the initial point is not sufficiently close to the solution (which is precisely what one seeks to calculate). Furthermore, J. Carrayrou limits the risk of overflow or under-filling by forcing the method to search for the solution in a neighborhood of a 'reasonable' value and he defines this reasonable neighborhood as an authorized chemical interval. In this work, we use the PCF method to reformulate the chemical problem as a fixed point problem.

Positive continuous fraction method PCF The continuous fraction method (CF) has been used to solve thermodynamic equilibrium in the computer code WATSPEC [51], or for preconditioning of the Newton-Raphson method for the major species in the PHREEQC code [33]. This method, which only needs one computation of the approximate thermodynamic equilibrium per iteration, is the cheapest zero-order method. Often, the component H^+ has a zero total concentration and is associated with negative stoichiometric coefficients. In the code WATSPEC [51], the pH value must be imposed to find the thermodynamic equilibrium. Hydrogen and oxygen are excluded from the continuous fraction preconditioning in the code PHREEQC [33]. Moreover, it has never been used for non ideal system.
To take into account a component with zero or negative total concentration, and to be more efficient with negative stoichiometric coefficients, a generalization of the (CF) method has been developed, called the positive continues fraction method (PCF) by Carrayrou [12]. This new method is an empirical method. Once the equilibrium solution is found, the reactive sum $\mathcal{S}^R$ is equal to the product sum $\mathcal{S}^P$. The *reactive sum* ($\mathcal{S}^R$) and the *product sum* ($\mathcal{S}^P$) are defined by

$$\mathcal{S}_j^R = \begin{cases} \sum\limits_{\mu_{i,j}>0} \mu_{i,j}.\mathcal{C}_i & \text{if } \mathbf{T}_j \geq 0 \\ |\mathbf{T}_j| + \sum\limits_{\mu_{i,j}>0} \mu_{i,j}.\mathcal{C}_i & \text{if } \mathbf{T}_j < 0 \end{cases} \text{ and } \mathcal{S}_j^P = \begin{cases} \mathbf{T}_j + \sum\limits_{\mu_{i,j}<0} |\mu_{i,j}|.\mathcal{C}_i & \text{if } \mathbf{T}_j \geq 0 \\ \sum\limits_{\mu_{i,j}<0} |\mu_{i,j}|.\mathcal{C}_i & \text{if } \mathbf{T}_j < 0 \end{cases} \tag{2.12}$$

for $j = 1, ..., n_p$, with $\mathcal{C} = \begin{pmatrix} C \\ CS \end{pmatrix}$ and $n_p = n_{pm} + n_{pf}$.
Using these two new values, the mass balance (2.7) is written, at equilibrium, as

$$\mathcal{S}_j^R = \mathcal{S}_j^P.$$

The coefficient $\mu_{i_0,j}$ is taken as the smallest value of the strictly positive stoichiometric coefficient in the matrix μ. The mass action laws are written for the reactive sum if $\mu_{i_0,j}$ is positive (respectively, for the product sum if $\mu_{i_0,j}$ is negative) by using component concentrations at iterations n and $n+1$. In particular, the following equality holds

$$(\mathcal{X}_j^{n+1})^{\mu_{i_0,j}}.\Big[\sum_{\mu_{i,j}>0} \mu_{i,j}\mathbf{K}_i \prod_{k\neq j} (\mathcal{X}_j^n)^{\mu_{i,k}}.(\mathcal{X}_j^n)^{\mu_{i,j}-\mu_{i_0,j}}\Big] = \mathbf{T}_j + \sum_{\mu_{i,j}<0} |\mu_{i,j}|\mathcal{C}_i^n, \tag{2.13}$$

where $\mathcal{X} = \begin{pmatrix} X \\ S \end{pmatrix}$, $\mathcal{X}_j^n$ is the concentration of the jth component species $\mathcal{X}_j$ at iteration n and $\mathcal{C}_i^n$ is that of the ith secondary species $\mathcal{C}_i$ at iteration n. After reordering, (2.13) becomes

$$(\mathcal{X}_j^{n+1})^{\mu_{i_0,j}} = \frac{(\mathcal{X}_j^n)^{\mu_{i_0,j}}}{(\mathcal{X}_j^n)^{\mu_{i_0,j}}} \frac{\mathbf{T}_j + \sum_{\mu_{i,j}<0} |\mu_{i,j}| \mathcal{C}_i^n}{\sum_{\mu_{i,j}>0} \mu_{i,j} \mathbf{K}_i \prod_{k\neq j} (\mathcal{X}_j^n)^{\mu_{i,k}} . (\mathcal{X}_j^n)^{\mu_{i,j}-\mu_{i_0,j}}}.$$

Then, the relationship (2.14) giving $\mathcal{X}_j^{n+1}$ is

$$\mathcal{X}_j^{n+1} = \mathcal{X}_j^n \left(\frac{\mathcal{S}_j^{P,n}}{\mathcal{S}_j^{R,n}}\right)^{\frac{1}{\mu_{i_0,j}}}. \tag{2.14}$$

Since $\omega^{n+1} = \log_{10}(\mathcal{X}^{n+1})$, then, written according to the logarithm of the component species concentrations, the relation (2.14) becomes

$$\omega_j^{n+1} = \omega_j^n + \frac{1}{\mu_{i_0,j}} \Big[\log_{10}(\mathcal{S}_j^{P,n}) - \log_{10}(\mathcal{S}_j^{R,n})\Big]. \tag{2.15}$$

This relation is considered to be the conventional fixed point iteration

$$\omega^{n+1} = \mathbf{G}(\omega^n),\ n = 0, 1, ..., \tag{2.16}$$

where $\mathbf{G} : \mathbb{R}^{n_p} \to \mathbb{R}^{n_p}$ is the fixed point map defined by

$$\mathbf{G}(\omega) = \omega + \frac{1}{\mu_0} \Big[\log_{10}(\mathcal{S}^P) - \log_{10}(\mathcal{S}^R)\Big]. \tag{2.17}$$

Thus, solving the chemical equilibrium problem (2.8) amounts to solving the fixed point problem

$$\omega = \mathbf{G}(\omega). \tag{2.18}$$

3. Iterative methods

The aim is to solve the previous nonlinear fixed point problem (2.18) whose the solution is denoted by ω^*. Then starting with a suitable vector ω_0, as an initial approximation to ω^*, the sequence $\{\omega_n\}$ is generated by fixed point iterative (FPI) methods defined by (2.16).

3.1. Anderson Acceleration

To improve the convergence rate of FPI (2.16), Anderson acceleration [3] is applied. It is formulated as follows [49]:

Algorithm 1: Anderson Acceleration (AA).

Given ω_0 *and* $m \geq 1$.
Set $x_1 = \mathbf{G}(\omega_0)$ *and* $f_0 = \mathbf{G}(\omega_0) - \omega_0$
For $k = 0, 1, ...$
Set $m_k = \min\{m, k\}$.
Compute $\mathbf{G}(\omega_k)$ *and let* $f(\omega_k) = \mathbf{G}(\omega_k) - \omega_k$.
Set $F_k = (f_{k-m_k}, ..., f_k)$.
Determine $\alpha_{(k)} = (\alpha_0^{(k)}, ..., \alpha_{m_k}^{(k)})^T$ *that solves*

$$\min_{\alpha=(\alpha_0,...,\alpha_{m_k})^T} ||F_k\alpha||_2,\ \text{s.t}\ \sum_{i=0}^{m_k} \alpha_i = 1 \tag{3.1}$$

Set $\omega_{k+1} = (1-\beta_k)\sum_{i=0}^{m_k} \alpha_i^{(k)} (\omega_{k-m_k+i}) + \beta_k \sum_{i=0}^{m_k} \alpha_i^{(k)} \mathbf{G}(\omega_{k-m_k+i})$

where $\beta_k > 0$ is a relaxation parameter. In [49], it is shown that Anderson acceleration with $\beta_k = 1$ converges when the fixed point map $\mathbf{G}$ is a contraction and that the rate of convergence is comparable to that of the Picard iteration. In this work, as in [49], only the case $\beta_k = 1$ in Algorithm 1 is considered. If $m = 0$, then Anderson acceleration becomes the FPI (2.16).

Form of least-squares problem In practical implementation, the constrained least-squares problem (3.1) is often formulated as the following equivalent unconstrained least-squares problem ([22], [49]):
Find $\gamma^{(k)} = (\gamma_0^{(k)}, ..., \gamma_{m_k-1}^{(k)})^T$ such that

$$\min_\gamma ||f_k - \mathcal{F}_k\gamma||_2 \tag{3.2}$$

where

$$\mathcal{F}_k = (\Delta f_{k-m_k}, ..., \Delta f_{k-1}) \tag{3.3}$$

with $\Delta f_i = f_{i+1} - f_i$ for $i = k - m_k, ..., k-1$. The least-squares coefficient vectors α and γ are related by $\alpha_0 = \gamma_0$, $\alpha_j = \gamma_j - \gamma_{j-1}$ for $1 \leq j \leq m_k - 1$ and $\alpha_{m_k} = 1 - \gamma_{m_k-1}$. The next iterate then becomes $\omega_{k+1} = \mathbf{G}(\omega_k) - \sum_{i=1}^{m_k-1} \gamma_i^{(k)}[\mathbf{G}(\omega_{k-m_k+i+1}) - \mathbf{G}(\omega_{k-m_k+i})] = \mathbf{G}(\omega_k) - \mathcal{G}_k\gamma^{(k)}$, where

$$\mathcal{G}_k = (\Delta\mathbf{G}_{k-m_k}, ..., \Delta\mathbf{G}_{k-1}) \tag{3.4}$$

with $\Delta\mathbf{G}_i = \mathbf{G}(\omega_{i+1}) - \mathbf{G}(\omega_i)$ for $i = k - m_k, ..., k-1$.
Then, a more specific version of the AA algorithm can be given in Algorithm 2.

Algorithm 2: ANDERSON ACCELERATION (AA).

Given ω_0 *and* $m \geq 1$.
Set $\omega_1 = \mathbf{G}(\omega_0)$ *and* $f_0 = \mathbf{G}(\omega_0) - \omega_0$.
For $k = 1, 2, ...$
  Set $m_k = \min(m, k)$.
  Compute $\mathbf{G}(\omega_k)$ *and let* $f_k = \mathbf{G}(\omega_k) - \omega_k$.
  Update $\mathcal{F}_k$ *and* $\mathcal{G}_k$ *by* (3.3) *and* (3.4).
  Determine $\gamma^{(k)} = (\gamma_0^{(k)}, ..., \gamma_{m_k-1}^{(k)})$ *that solves* $\min_{(\gamma_0,...,\gamma_{m_k-1})^T} ||f_k - \mathcal{F}_k\gamma||_2$.
  Set $\omega_{k+1} = \mathbf{G}(\omega_k) - \mathcal{G}_k\gamma^{(k)}$.

The least-squares problem (3.2) is solved by performing the QR factorization of $\mathcal{F}_k$ and using backward substitution to solve the upper triangular system $R_k\gamma = Q_k^T f_k$. This shows that only Q_k and R_k need to be computed. In other words, only the "thin" QR decomposition of $\mathcal{F}_k$ ($\mathcal{F}_k = Q_k R_k$, $Q_k \in \mathbb{R}^{N\times m_k}$ and $R_k \in \mathbb{R}^{m_k\times m_k}$) is necessary.
Since $\mathcal{F}_k$ is obtained from $\mathcal{F}_{k-1}$ by appending a new column on the right and possibly dropping one column from the left, the QR decomposition of $\mathcal{F}_k$ can be efficiently obtained by updating that of $\mathcal{F}_{k-1}$. For details about this aspect, see [48].

Condition control In practice, there is often a risk that $\mathcal{F}_k$ will become ill-conditioned as iterations go. In this work, the strategy given in [49] is used to monitor the condition number of the matrix $\mathcal{F}_k$ and, if necessary, to modify the matrix to reduce the condition number, as follows: when the condition number of $\mathcal{F}_k$ is larger than a given tolerance, then the left-most columns of $\mathcal{F}_k$ are dropped one by one until the condition number is less than the given tolerance. Note that the l^2-norm condition number of $\mathcal{F}_k$ is just that of R_k in the QR decomposition of $\mathcal{F}_k$. Therefore, for the filtering strategy used in this paper, it is only necessary to monitor the condition number of R_k and keep it less than the given tolerance. If the condition number of R_k is larger than the given tolerance, then removing the leftmost column of $\mathcal{F}_k$ involves updating the factors Q_k and R_k (see [48] for details).

3.2. POLYNOMIAL VECTOR EXTRAPOLATION METHODS MPE AND RRE

An important problem that arises in different areas of science and engineering is that of computing the limits of sequences of vectors. Vector sequence arises, for example, in the solution of system of linear or nonlinear equations by fixed-point iterative methods, its limit being simply the required solution.
Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^N$, and define the first and second forward differences such that

$$\Delta x_k = x_{k+1} - x_k \quad \text{and} \quad \Delta^2 x_k = \Delta x_{k+1} - \Delta x_k \quad k = 0, 1, ...$$

When MPE and RRE are applied to the vector sequence $\{x_k\}$, an approximation $t_{k,q}$ is produced of the limit or antilimit of $\{x_k\}_{k\in\mathbb{N}}$ (cf.[41]). It is clear that t_k will be different for each method. Let

$$t_{k,q} = \sum_{j=0}^{k} \nu_j^{(k)} x_{q+j} \tag{3.5}$$

subject to

$$\sum_{j=0}^{k}\nu_j^{(k)}=1 \quad \text{and} \quad \sum_{j=0}^{k}\tau_{i,j}\nu_j^{(k)}=0 \quad i=0,1,...,k-1 \tag{3.6}$$

with the scalars $\tau_{i,j}$ defined by the inner products in $\mathbb{R}^N$: $\tau_{i,j}=\begin{cases}(\Delta x_{q+i},\Delta x_{q+j}) & \text{for MPE}\\ (\Delta^2 x_{q+i},\Delta x_{q+j}) & \text{for RRE}\end{cases}$.
Using (3.6), the transformation (3.5) can also be expressed as a ratio of two determinants as follows

$$t_{k,q}=\frac{\begin{vmatrix} x_q & x_{q+1} & \dots & x_{q+k}\\ \tau_{0,0} & \tau_{0,1} & \dots & \tau_{0,k}\\ \vdots & \vdots & & \vdots\\ \tau_{k-1,0} & \tau_{k-1,1} & \dots & \tau_{k-1,k}\end{vmatrix}}{\begin{vmatrix} 1 & 1 & \dots & 1\\ \tau_{0,0} & \tau_{0,1} & \dots & \tau_{0,k}\\ \vdots & \vdots & & \vdots\\ \tau_{k-1,0} & \tau_{k-1,1} & \dots & \tau_{k-1,k}\end{vmatrix}}.$$

Matrices $\Delta^i S_{k,q}=[\Delta^i x_q,...,\Delta^i x_{q+k-1}]$, $i=1,2$, are introduced. Using Schur complements, $t_{k,q}$ can be written, for each method as

$$t_{k,q}^{\text{MPE}}=x_q-\Delta S_{k,q}(\Delta S_{k,q}^T\Delta^2 S_{k,q})^{-1}\Delta S_{k,q}^T\Delta x_q$$
$$t_{k,q}^{\text{RRE}}=x_q-\Delta S_{k,q}(\Delta^2 S_{k,q}^T\Delta^2 S_{k,q})^{-1}\Delta^2 S_{k,q}^T\Delta x_q$$

provided that $\det(\Delta S_{k,q}^T\Delta^2 S_{k,q})\neq 0$ and $\det(\Delta^2 S_{k,q}^T\Delta^2 S_{k,q})\neq 0$. These two assumptions are assumed in the following. Then $t_{k,q}^{\text{MPE}}$ and $t_{k,q}^{\text{RRE}}$ are well defined and unique. For varying value of k and q, the computation of $t_{k,q}^{\text{MPE}}$ and $t_{k,q}^{\text{RRE}}$ can be done by some of algorithms proposed by Ford and Sidi in [42].

An estimate for the residual norm for nonlinear problems is given. Introduce the new approximation

$$\tilde{t}_{k,q}=\sum_{j=0}^{k}\nu_j^{(k)}x_{q+j+1}$$

In [25], the generalized residual of $t_{k,q}$ as is defined by

$$\tilde{r}(t_{k,q})=\tilde{t}_{k,q}-t_{k,q}, \tag{3.7}$$

which can be expressed as

$$\tilde{r}(t_{k,q}^{\text{MPE}})=\Delta x_q-\Delta^2 S_{k,q}(\Delta S_{k,q}^T\Delta^2 S_{k,q})^{-1}\Delta S_{k,q}^T\Delta x_q$$
$$\tilde{r}(t_{k,q}^{\text{RRE}})=\Delta x_q-\Delta^2 S_{k,q}(\Delta^2 S_{k,q}^T\Delta^2 S_{k,q})^{-1}\Delta^2 S_{k,q}^T\Delta x_q.$$

Implementation Only the case q kept constant is considered. Without restriction, $q=0$ is always assumed and $t_{k,0}$ is denoted by t_k; $\Delta^i S_{k,0}$ by $\Delta^i S_k$. The linear system (3.6) can be written as

$$\begin{array}{ccccccccc}
\nu_0^{(k)} & + & \nu_1^{(k)} & + & \dots & + & \nu_k^{(k)} & = & 1\\
\nu_0^{(k)}(u_0,\Delta x_0) & + & \nu_1^{(k)}(u_0,\Delta x_1) & + & \dots & + & \nu_k^{(k)}(u_0,\Delta x_k) & = & 0\\
\nu_0^{(k)}(u_1,\Delta x_0) & + & \nu_1^{(k)}(u_1,\Delta x_1) & + & \dots & + & \nu_k^{(k)}(u_1,\Delta x_k) & = & 0\\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots\\
\nu_0^{(k)}(u_{k-1},\Delta x_0) & + & \nu_1^{(k)}(u_{k-1},\Delta x_1) & + & \dots & + & \nu_k^{(k)}(u_{k-1},\Delta x_k) & = & 0
\end{array} \tag{3.8}$$

Introduce the scalars $\theta_i^{(k)}=\frac{\nu_i^{(k)}}{\nu_k^{(k)}}$, for $i=0,...,k$. Then, $\nu_i^{(k)}=\frac{\theta_i^{(k)}}{\sum_{i=0}^{k}\theta_i^{(k)}}$, for $i=0,...,k-1$, and $\theta_k^{(k)}=1$.
With this new variables, the linear system (3.8) becomes

$$\begin{array}{ccccccccc}
\theta_0^{(k)}(u_0,\Delta x_0) & + & \theta_1^{(k)}(u_0,\Delta x_1) & + & \dots & + & \theta_{k-1}^{(k)}(u_0,\Delta x_{k-1}) & = & -(u_0,\Delta x_k)\\
\vdots & & \vdots & & & & \vdots & & \vdots\\
\theta_0^{(k)}(u_{k-1},\Delta x_0) & + & \theta_1^{(k)}(u_{k-1},\Delta x_1) & + & \dots & + & \theta_{k-1}^{(k)}(u_{k-1},\Delta x_{k-1}) & = & -(u_{k-1},\Delta x_k)
\end{array}$$

This system can be written in the following form

$$(U_k^T \Delta S_k)\theta^{(k)} = -U_k^T \Delta x_k \tag{3.9}$$

where $\theta^{(k)} = (\theta_0^{(k)}, \ldots, \theta_{k-1}^{(k)})^T$, $\Delta S_k = (\Delta x_0, ..., \Delta x_{k-1})$ and $U_k = \begin{cases} \Delta S_k & \text{for the MPE method} \\ \Delta^2 S_k & \text{for the RRE method.} \end{cases}$

Assume now that the coefficients $\nu_0^{(k)}, ..., \nu_k^{(k)}$ have been calculated and introduce the new variables

$$\sigma_0^{(k)} = 1 - \nu_0^{(k)}, \quad \sigma_j^{(k)} = \sigma_{j-1}^{(k)} - \nu_j^{(k)},\ j = 1, ..., k-1, \quad \text{and } \sigma_{k-1}^{(k)} = \nu_k^{(k)}.$$

Then, for both method, the vector t_k can be expressed as

$$t_k = x_0 + \sum_{j=0}^{k-1} \sigma_j^{(k)} \Delta x_j = x_0 + \Delta S_k \sigma^{(k)} \tag{3.10}$$

where $\sigma = (\sigma_0, ..., \sigma_{k-1})^T$.

Note that to determine the coefficient $\nu_i^{(k)}$, we must first calculate the $\theta_i^{(k)}$ by solving the linear system of equations (3.9). Using (3.7) and (3.10), the generalized residual $\tilde{r}(t_k)$, for MPE and RRE, can be expressed as

$$\tilde{r}(t_k) = \sum_{i=0}^{k} \nu_i^{(k)} \Delta x_i = \Delta S_{k+1} \nu^{(k)}.$$

Algorithms for RRE and MPE methods Fast, stable, and storage wise economical algorithms are described in [24]. These algorithms solve least-squares problems by QR factorization. An overview of these algorithms is provided in the following.
ΔS_{k+1} has a full rank, namely $\mathbf{rank}(\Delta S_{k+1}) = k+1$. Then a QR factorization of ΔS_{k+1} can be computed.
For RRE method, this QR decomposition is defined by $\Delta S_{k+1} = Q_k R_K$ where $Q_k = (q_0|q_1|...|q_k) \in \mathbb{R}^{N\times(k+1)}$ has orthonormal columns q_j and $R_k \in \mathbb{R}^{(k+1)\times(k+1)}$ is an upper triangular matrix with positive diagonal coefficients. Q_k is obtained from $Q_{k-1} \in \mathbb{R}^{N\times k}$ by adding the column q_k. In the same way, R_k is obtained from $R_{k-1} \in \mathbb{R}^{k\times k}$ by adding a row and a column to R_{k-1}.
For MPE method, $\Delta S_{k+1} = Q_{k+1} R_{k+1}$, where $Q_{k+1} = (q_0|q_1|...|q_k) \in \mathbb{R}^{N\times(k+1)}$ is an orthogonal matrix and $R_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ is an upper triangular matrix with positive diagonal coefficients. Q_{k+1} is obtained from $Q_k \in \mathbb{R}^{N\times k}$ by adding the vector column q_k. Similarly, R_{k+1} is obtained from $R_k \in \mathbb{R}^{k\times k}$ by adding a row and a column to R_k.
For both method, the QR factorization of ΔS_{k+1} can be computed inexpensively by applying the modified Gram-Schmidt process (MGS) to the vectors $x_0, x_1, ..., x_{k+1}$ (cf. MGS algorithm in [43]).
The details of previous algorithms for RRE (resp. MPE) are summarized in Algorithm 3 (resp. Algorithm 4). Note that, in these algorithms, it is only necessary to store the vector x_0 and the matrix Q_k. The rest can be overwritten as soon as they have been used.

Algorithm 3: RRE METHOD

0. *Inputs: Vectors* $x_0, x_1, ..., x_{k+1}$.
1. *Compute* $v_i = \Delta x_i = x_{i+1} - x_i,\ i = 0, 1, ..., k$.
Set $V_j = [v_0|v_1|...|v_{j-1}],\ j = 0, 1, ...$
Compute the QR factorization of V_{k+1}, *namely* $V_{k+1} = Q_k R_k$.
($V_k = Q_{k-1} R_{k-1}$ *is contained in* $V_{k+1} = Q_k R_k$).
2. *Computation of the* ν_i:
Solve the linear system: $R_k^T R_k d^{(k)} = e$; $d^{(k)} = [d_0^{(k)}, d_1^{(k)}, ..., d_k^{(k)}]^T$; $e = [1, 1, ..., 1]^T$.
(This amounts to solving two upper and lower triangular systems).
Set $\lambda = (\sum_{i=0}^{k} d_i^{(k)})^{-1},\ \lambda \in \mathbb{R}^+$.
Set $\nu_i^{(k)} = \lambda d_i^{(k)},\ i = 0, 1, ..., k$.
3. *Compute* $\sigma^{(k)} = [\sigma_0^{(k)}, \sigma_1^{(k)}, ..., \sigma_{k-1}^{(k)}]^T$ *by:* $\sigma_0^{(k)} = 1 - \nu_0^{(k)}$ *and* $\sigma_j^{(k)} = \sigma_{j-1}^{(k)} - \nu_j^{(k)},\ j = 1, ..., k-1$.
Compute t_k *via:* $t_k^{\text{RRE}} = x_0 + Q_{k-1}(R_{k-1}\sigma^{(k)})$

Algorithm 4: MPE METHOD

0. *Inputs: Vectors* $x_0, x_1, ..., x_{k+1}$.
1. *Compute* $v_i = \Delta x_i = x_{i+1} - x_i,\ i = 0, 1, ..., k$.
Set $V_j = [v_0|v_1|...|v_j],\ j = 0, 1, ...$

Compute the QR factorization of V_{k+1}, *namely* $V_{k+1} = Q_{k+1}R_{k+1}$.
($V_k = Q_kR_k$ *is contained in* $V_{k+1} = Q_{k+1}R_{k+1}$).

2. *Computation of the* ν_i:
Solve the upper triangular linear system: $R_k d^{(k)} = -r_k$; $d^{(k)} = [d_0^{(k)}, ..., d_{k-1}^{(k)}]^T$; $r_k = [r_{0k}, ...r_{(k-1)k}]^T$.
Set $d_k^{(k)} = 1$ and calculate $\lambda = (\sum_{i=0}^{k} d_i^{(k)})^{-1}$, $\lambda \in \mathbb{R}^+$.
Set $\nu_i^{(k)} = \lambda d_i^{(k)}$, $i = 0, 1, ..., k$.

3. *Compute* $\sigma^{(k)} = [\sigma_0^{(k)}, \sigma_1^{(k)}, ..., \sigma_{k-1}^{(k)}]^T$ *by:* $\sigma_0^{(k)} = 1 - \nu_0^{(k)}$ *and* $\sigma_j^{(k)} = \sigma_{j-1}^{(k)} - \nu_j^{(k)}$, $j = 1, ..., k-1$.
Compute t_k *via:* $t_k^{\text{MPE}} = x_0 + Q_k(R_k\sigma^{(k)})$.

Algorithms 3 and 4 become increasingly expensive as the number of iteration steps k is increasing. Indeed the work requirement grows quadratically with k and the storage requirement grows linearly. A good way to keep the storage requirement and the computation cost low is to periodically restart the RRE and MPE algorithms every c steps, for some integer $c > 1$. Below, a practical strategy of a restarted method is described in Algorithm 5.

Algorithm 5: CYCLIC METHOD EVERY c ITERATIONS

For $k = 0$, *choose an integer* c *and an initial vector* x_0.
For $k = 1, 2, ...,$
Compute the vectors $x_1, ..., x_c$.
Calculate t_{c-1} *using the algorithm of the desired method*.
If t_{c-1} *satisfies accuracy test, stop*;
Else, set $x_0 = t_{c-1}$.

Similarly to linear problems [43], it is more useful to run some basic iterations before applying one of the extrapolation methods for solving (2.18):

- Let run some N_0 basic iterations before cycling is started, e.g, before MPE or RRE is applied for the first time (N_0 refers to the size of extrapolation);
- Let run some N basic iterations before MPE or RRE is applied in each cycle after the first cycle.

One way to make the extrapolation process more efficient with high numerical stability is to change (2.16) as follows

$$\omega_{n+1} = \tilde{\mathbf{G}}(\omega_n) \quad n = 0, 1, ..., \tag{3.11}$$

where

$$\tilde{\mathbf{G}}(\omega) = \omega + \kappa(\mathbf{G}(\omega) - \omega). \tag{3.12}$$

The scalar κ is different than 1 (the sequence generated by taking $\kappa = 1$ is the one generated by (2.16)). Thus ω_{n+1} is now weighted "average" of ω_n and $\mathbf{G}(\omega_n)$, in which weights $1 - \kappa$ and κ do not need to both be positive. By picking κ appropriately, the spectrum of the Jacobian matrix of $(1 - \kappa)\omega + \kappa\mathbf{G}(\omega)$ at $\omega = \omega^*$ can be taken as it is increasingly favorable to $t_{k,q}$ for large values of q ([43]; Section 7). So, with ω_0, the initial approximation of ω^*, we generate the sequence $\omega_1, \omega_2, \omega_3, ...$ by the fixed-point iteration (3.11). We consider the following algorithm:

Algorithm 6: EXTRAPOLATION ALGORITHM FOR THE NONLINEAR SYSTEM (2.18)

1. *For* $k = 0$, *choose* ω_0 *and the integers* p *and* l.

2. Basic iteration:
Set $t_0 = \omega_0$.
$h_0 = t_0$.
$h_{j+1} = \tilde{\mathbf{G}}(h_j)$, $j = 0, ..., p-1$.

3. Extrapolation phase:
$s_0 = h_p$.
If $||s_1 - s_0|| < \epsilon$, *stop*,
Else $s_{j+1} = \tilde{\mathbf{G}}(s_j)$, $j = 0, ..., l$.
Compute the approximation t_l *by* RRE *or* MPE.

4. *Set* $\omega_0 = t_l$, $k = k + 1$ *and go to* **2.**

4. NUMERICAL EXPERIMENTS

In this section, numerical experiments are reported. Iterative numerical methods cited above are tested to resolve the chemical equilibrium of two different chemical systems. Anderson acceleration is implemented using

the approach cited in section 3.1, in its specific version (Algorithms 1-2). Sometimes, (AA) method is applied to the fixed point problem with relaxation (3.12) instead of (2.18), with $\kappa \neq 1$. The two vector extrapolation methods of the polynomial type MPE and RRE are also applied to the problem of the nonlinear chemical equilibrium system thanks to the implementations previously described. These implementations are done by employing the computer program which is provided in [43] using Matlab R2018a (this computer program written in Fortran has been converted in Matlab language). It gives an estimation of the residual norm at each iteration for this nonlinear problem and it allows to stop algorithms without having to compute the true residual which requires an extra evaluation of $\tilde{\mathbf{G}}$.

Some essential numerical parameters for test cases are:
For Anderson Acceleration:

- the iteration is stopped when the residual norm falls below 10^{-10};
- the condition-number monitoring is used with a threshold for deleting columns $droptol = 10^{10}$;
- The allowed maximal nonlinear iteration number is $Kmax = 200$ iterations.

For MPE and RRE:

- the maximum number of cycles allowed is $Ncycle = 30$;
- the upper bound of $resc/resp$ used in the stopping criterion is $epsc = 10^{-10}$ (where $resp$ is the l_2-norm of the residual for $t(Kmax, N_0)$ at the end of the first cycle and $resc$ is the l_2-norm of the residual at t at the end of each cycle, retrieved at the end of the next cycle). If $resc \leq epsc \times resp$ at the end of some cycle, then one additional cycle is performed, and the corresponding $t(N, Kmax)$ is accepted as the final approximation.
- the upper bound of $res/R(0,0)$, the relative residual for t, used in the stopping criterion is $eps = 0$. Note that $R(0,0) = l_2$ norm of the residual of ω_0, the initial vector. If , for some k, $res \leq eps \times R(0,0)$ then the corresponding $t(0,k)$ is accepted as the final approximation.

4.1. Chemical tests

Gallic acid test This is the simplest test, a system proposed by Brassard and Bodurtha (2000) [6] to illustrate the appearance of problems with numerical methods. The system was originally studied for the speciation of Al(III) in natural waters. It is characterized by the presence of $n_e = 17$ chemical species that can be described through the combination of $n_{pm} = 3$ mobile components species ($n_{sm} = 14$). All reactions describing this chemical system are homogeneous, between the mobile species ($n_{pf} = n_{sf} = 0$). The pH is imposed at 5.8, which gives a problem with two unknowns: concentrations of the two free components Al^{3+} and H_3L.
The chemical system studied is presented in Table 1 where the initial concentrations of Al^{3+} and H_3L are variable and thermodynamic values come from Brassard and Bodurtha (2000) [6]. By fixing the pH of the system, we notice that the matrix of the stoichiometric coefficients is reduced to a matrix whose all the coefficients are positive, moreover, total concentrations of Al^{3+} and H_3L are positive, then the relation (3.12) becomes, in matrix form

$$\tilde{\mathbf{G}}(\xi) = \xi + \kappa \frac{I_{n_{pm}}}{\mu_{10}} . \Big[\log_{10}(T^m) - \log_{10}(\mu_1^T . 10^{\mathbf{K}^m + \mu_1 \xi}) \Big]. \tag{4.1}$$

Methods are tested in 2 cases:
case 1 - Initial X_0 (M): $[Al^{3+}]_0 = 10^{-11}$ M; $[H_3L]_0 = 5 \times 10^{-4}$ M
(i.e Initial ξ_0 : $\log_{10}([Al^{3+}]_0) = -11; \log_{10}([H3L]_0) = -3.3010$)
case 2 - Initial X_0 (M): $[Al^{3+}]_0 = 5.012 \times 10^{-10}$ M; $[H_3L]_0 = 10^{-9}$ M.
(i.e Initial ξ_0 : $\log_{10}([Al^{3+}]_0) = -9.3; \log_{10}([H3L]_0) = -9$)

MoMas Benchmark easy test case The MoMaS Benchmark has been designed to compare numerical methods for reactive transport model in 1D and 2D. Different methods for coupling have been used to solve this benchmark. The definition has been published in [13] and results of participants are compared in the synthesis article [14]. It is composed of three subsequent cases with increasing chemical complexity, named "easy", "medium" and "hard". Systems do not represent real chemical systems: they were devised by [5] to create increasing numerical difficulties.
In this work, only the resolution of the chemical equilibrium of the easy test case will be simulated. For the easy case, the chemical system is composed of $n_e = 12$ chemical species distributed as follows: four mobile component species X_1, X_2, X_3 and X_4 ($n_{pm} = 4$), one fixed component species S ($n_{pf} = 1$), five mobile secondary species C_1, C_2, C_3, C_4 and C_5 ($n_{sm} = 5$), and two fixed secondary species CS_1 and CS_2 ($n_{sf} = 2$).
The geometry of the test case is shown in Figure 1. For the 1D test case, the domain is heterogeneous and

Species	H^+	Al^{3+}	H_3L	$\mathbf{K}^m$	C **(Equil.)**
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_3^{6-}	-9	1	3	-37.69	1.14×10^{-10}
$Al_2(OH)_2(HL)_3^{2-}$	-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}
$Al_3(OH)_4(H_2L)^{4+}$	-5	3	1	-12.52	6.51×10^{-5}
Total T^m (M)	pH=5.8	10^{-3}	10^{-3}		
Initial X_0 (M)	1.58×10^{-6}	variable	variable		
Equil. X^* (M)	1.58×10^{-6}	2.03×10^{-5}	2.59×10^{-7}		

TABLE 1. Morel's Table for the Gallic Acid Test with pH Fixed to 5.8 ([6], [12])

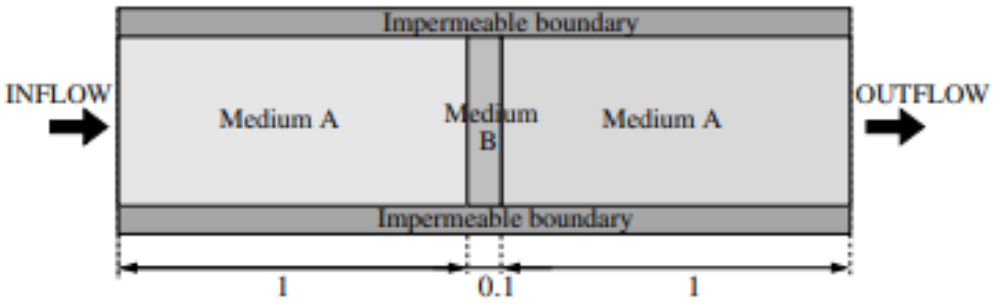

FIGURE 1. Geometry of the domain.

composed of two porous media A and B. Medium A is highly permeable with low porosity and low reactivity in comparison with medium B. In order to be close to realistic cases, boundary and initial conditions are not expressed for fundamental variables, i.e., component concentrations. Indeed, chemical analysis can provide quite easily a measure of the total concentration or of the total dissolved concentration for each component. An injection is made on the left side of the domain, followed by leaching on the same side. The injection period corresponds to specific inflow concentrations depending on the MoMas easy test case. All injection periods are 5000 s long. The leaching periods are at least 1000 s long. If needed, leaching period can be extended after 1000 s to reach the following condition: at the end of leaching period, $99,9\%$ of injected pollutant (X_1, X_3 and S) has been removed from the domain. Imposed concentrations for the inflow boundary are

$$T_j(x=0,t) = T_j^{inj} \quad t < 5000\ s \qquad T_j(x=0,t) = T_j^{leach} \quad t > 5000\ s$$

These chemical species interact through $n_r = 7$ equilibrium reactions shown in Table 2.
The domain is initially at a local equilibrium with the surface component S in the presence of mobile components X_2 and X_4. During injection, component X_4 will be removed. Component X_1 is a perfect tracer; X_2 and X_3 will react together, with the surface S and with X_4 still present. During leaching, X_1 and X_3 will be removed. X_2 and X_4 will react with the surface S and with X_3 still present.
In this test case, the equilibrium reaction constant is of the order of 10^{35} which makes the system very rigid and already presents a great challenge. In addition, stoichiometric coefficients are quite large. This allows us to test the robustness of our implementation in the face of such complexity. Table 3 give the stoichiometric

$$
\begin{aligned}
-X_2 &\rightleftarrows C_1 && (1)\\
X_2+X_3 &\rightleftarrows C_2 && (2)\\
-X_2+X_4 &\rightleftarrows C_3 && (3)\\
-4X_2+X_3+3X_4 &\rightleftarrows C_4 && (4)\\
4X_2+3X_3+X_4 &\rightleftarrows C_5 && (5)\\
3X_2+X_3+S &\rightleftarrows CS_1 && (6)\\
-3X_2+X_4+2S &\rightleftarrows CS_2 && (7)
\end{aligned}
$$

TABLE 2. Chemical reactions for MoMas easy test case

Species	X_1	X_2	X_3	X_4	S	**K**
C_1	0	-1	0	0	0	10^{-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	10^{35}
CS_1	0	3	1	0	1	10^{6}
CS_2	0	-3	0	1	2	10^{-1}
Total concentration T	T_1	T_2	T_3	T_4	T_S	
Initial conditions						
zone A	0	-2	0	2	1	
zone B	0	-2	0	2	10	
Boundary conditions						
Injection $t \in [0, 5000]$	0.3	0.3	0.3	0	0	
Leaching $t \in [5000, ...]$	0	-2	0	2	0	

TABLE 3. Equilibrium for MoMas easy test case [13]

coefficients for mass action laws and conservation equations.
The resolution of the thermodynamic equilibrium of this test is carried out at each period and results obtained correspond well to expected results. But first of all it is necessary to put the two medium A and B at local equilibrium, then the chemical equilibrium during injection and leaching periods will be resolved.

4.2. NUMERICAL RESULTS AND COMPARISONS

Acid Gallic test The thermodynamic chemical equilibrium problem (4.1) is studied without relaxation ($\kappa = 1$). For each case defined above, we notice the convergence of Anderson Acceleration method, for a maximal depth $m \geq 1$ (cf. Figure 2):

- for $\omega_0 = \log_{10}((10^{-11}, 5 \times 10^{-4})^T)$, i.e in case 1, the convergence of Anderson(m) requires 26 iterations for $m = 1$ and 16 iterations for $m \geq 2$.
- for $\omega_0 = \log_{10}((5.012 \times 10^{-10}, 10^{-9})^T)$, i.e in case 2, the convergence of Anderson(m) requires 109 iterations for $m = 1$ and 15 iterations for $m \geq 2$.

The first iterations performed present disturbances in terms of the variation in concentrations $[Al^{3+}]$ et $[H_3L]$, but these disturbances are no complicated and convergence has been obtained without any difficulty. These disturbances result mainly from the choice of the initial concentration of each component. Note that theobtained solution $\omega^* = (-4.6930, -6.5870)^T = \log_{10}((2.028 \times 10^{-5}, 2.6 \times 10^{-7})^T)$ is the same reference solution obtained by J. Carrayrou [12] and cited in Table 1. It is thus numerically established that Anderson Acceler-

	CPU time (s)
Anderson $(m = 1)$	1.11
Anderson $(m = 2)$	1.3
Anderson $(m = 3)$	1.13

TABLE 4. Gallic acid test, thermodynamic equilibrium by Anderson Acceleration method: CPU time (s)

ation method converges towards the solution in a short computation time (CPU time) (cf.Table 4) for different

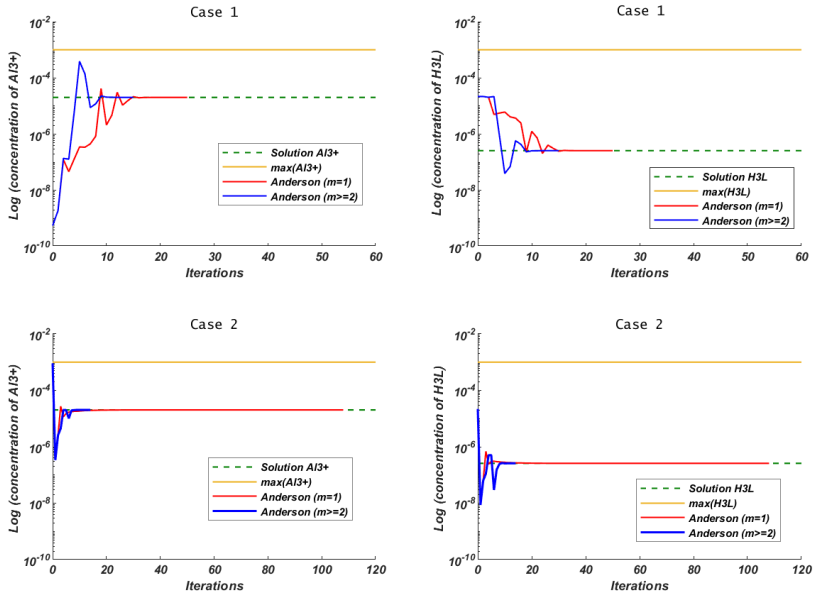

FIGURE 2. Acid Gallic test: Thermodynamic equilibrium for the components H_3L and Al^{3+} with Anderson Acceleration

values of the maximal depth m, $m = 1, 2, 3$.
Since a condition-number monitoring strategy is used in Anderson acceleration, it is not necessary to worry about the condition number becoming problematically large. In this Anderson-acceleration implementation, the condition number is monitored to ensure stability and robustness. The tolerance for the condition-number monitoring is 10^{10}. We can see in Figure 3 that, for $m = 1, 2$, the condition number remains less than 10^{10} and it becomes more than 10^{15} for $m \geq 3$ if there is no condition-number monitoring. This shows the specific effects of condition-number monitoring. With condition-number monitoring, $\mathbf{cond}(\mathcal{F}_3)$, at iteration step 3, is

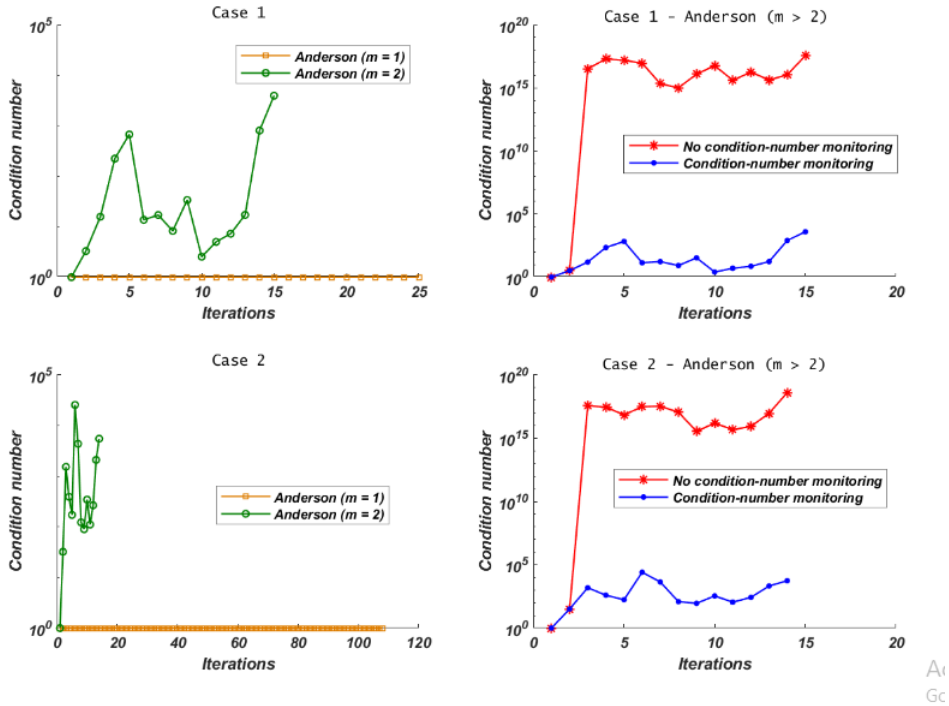

FIGURE 3. Acid Gallic test: Thermodynamic equilibrium by Anderson Acceleration - Condition number curve.

initially greater than 10^{10} ($\mathcal{F}_k$ is defined by (3.3)). However, after using the MATLAB's qrdelete function, the

condition number is less than 10^{10}, and the convergence succeeds after 3 iterations.
In addition, from Figure 4 we can see that for $m \geq 2$, Anderson Acceleration method begins to accelerate the

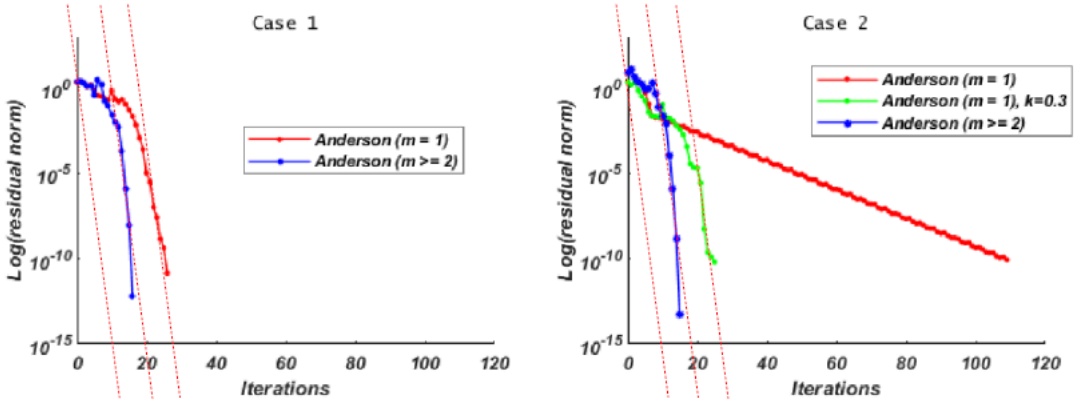

FIGURE 4. Acid Gallic test: Thermodynamic equilibrium by Anderson Acceleration - Residual norm curve.

convergence of the AA iterates after 12 or 13 iterations in the both case, but it is clear that results slowly decrease when $m = 1$ especially in the second case where convergence requires more than 100 iterations. More precisely, if we now take $\kappa = 0.3$, Anderson ($m = 1$) converges faster and the residual norm decreases requiring 26 iterations instead of 109, to be less than 10^{-10}. On the other hand, the numerical slopes of the tangents to the curve are plotted on the Figure 4. These numerical slopes show that a convergence of order 2 is reached, which confirms that the AA method works really well.

To apply the two methods MPE and RRE, we generate the vectors $\omega_1, \omega_2, \ldots$ by (3.11) with different values of the parameter κ, $\kappa \in \{0.1, 0.45, 0.5, 0.6\}$. These values of κ are a good choice for this experience, but that does not mean that these parameters cannot take another value
Figures 5 and 6 show the behavior of the residual norm, using a logarithmic scale. It contains some of the residual history obtained by applying RRE and MPE in cycling mode with

- $(Kmax, N_0, N) = (10, 20, 10), (20, 0, 10), (10, 20, 0)$ for $\kappa = 0.1$,
- $(Kmax, N_0, N) = (10, 10, 15), (10, 10, 10)$ for $\kappa = 0.45$,
- $(Kmax, N_0, N) = (10, 5, 15), (20, 5, 15)$ for $\kappa = 0.5$,
- $(Kmax, N_0, N) = (15, 15, 15)$ for $\kappa = 0.6$.

Convergence behavior is overall linear, but a small marginally unstable mode is observed that corresponds to almost a periodic half-sinusoidal oscillation of residuals. With $\kappa = 0.1$, the first choice of the three data is the best because it yields the fastest convergence of the residual error. By comparing the first and third choices of the data $Kmax, N_0$ and N, we notice that performing a few basic iterations before MPE or RRE is applied in each cycle after the first cycle results in a faster and more stable convergence as well as more stable behavior of the residual norm. Such a result is observed again in Figure 6 with $\kappa = 0.45$, $\kappa = 0.5$ and $\kappa = 0.6$. These last three values of κ still give a fast convergence represented by an almost stable decrease of the residual norm.
In addition, the CPU time required for all iterations in all cycles, for each case and each method, is very short, not exceeding 1 second (see Table 5 for case 1). Therefore, for this chemical test, solving the thermodynamic

	$(Kmax, N_0, N)$	**MPE**	**RRE**
$\kappa = 0.1$	(10,20,10)	0.6406	0.8594
	(20,0,10)	0.6719	0.4688
	(10,20,0)	0.8594	0.3906
$\kappa = 0.45$	(10,10,10)	0.2656	0.2031
	(10,10,15)	0.4531	0.2969
$\kappa = 0.5$	(10,5,15)	0.2656	0.2656
	(20,5,15)	0.3906	0.5

TABLE 5. Acid Gallic test, thermodynamic equilibrium by restarted RRE and MPE: CPU time (s)

equilibrium using Anderson's acceleration method and the two polynomial methods MPE and RRE works

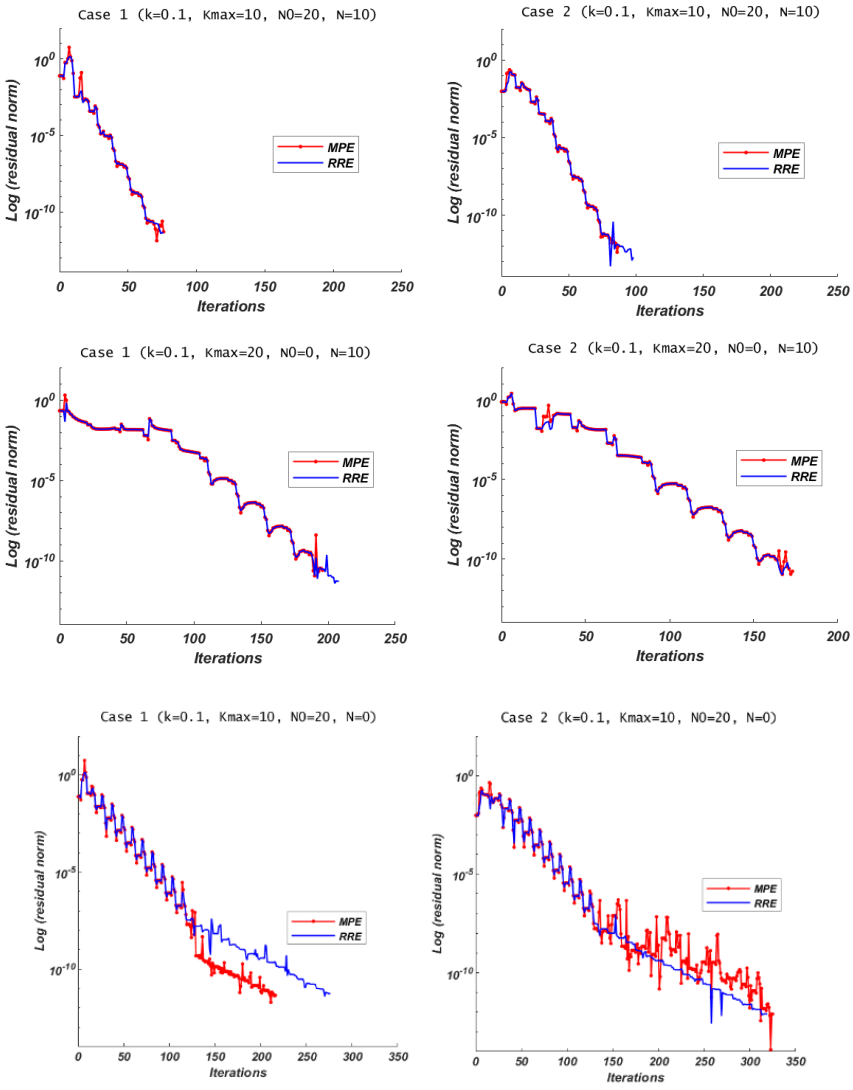

FIGURE 5. Acid Gallic test: Thermodynamic equilibrium by restarted MPE and RRE - Residual norm curve

well. The convergence is obtained in a very short computation time and a reasonable number of iterations, but Anderson's method seems to be more stable than the MPE and RRE methods, and requires fewer iterations. This results from the cycling strategy applied to the MPE and RRE methods, while the Anderson Acceleration method does not require such a strategy.
After having successfully calculated the equilibrium concentrations of the component species H_3L and Al^{3+} by the methods (AA), (MPE) and (RRE), we can then calculate the concentrations of equilibrium of the other secondary species from equations in (2.9). Equilibrium concentrations are reported in the last column of Table 1.

Benchmark MoMas easy test case In this part, numerical results for the resolution of the thermodynamic equilibrium of the easy MoMas test case are presented. First the chemical equilibrium in each medium (A and B) is solved, then the chemical equilibrium for the injection period of the component X_3, before looking for the chemical equilibrium of the leaching period.
We assume, for example, that the initial concentrations of the component species in each of the two mediums A and B are given by the vector $\omega_{A,B,0} = \log_{10}(\mathcal{X}_{A,B,0})$ where

$$\mathcal{X}_{A,B,0}\,(M) = (X_{1,0}, X_{2,0}, X_{3,0}, X_{4.0}, S_0)^T = (0.3, 0.4, 10^{-11}, 0.21, 0.6)^T.$$

The convergence of Anderson acceleration method is tested for any strictly positive value of the maximum depth m. The fixed point problem (3.11) is implemented without relaxation ($\kappa = 1$).

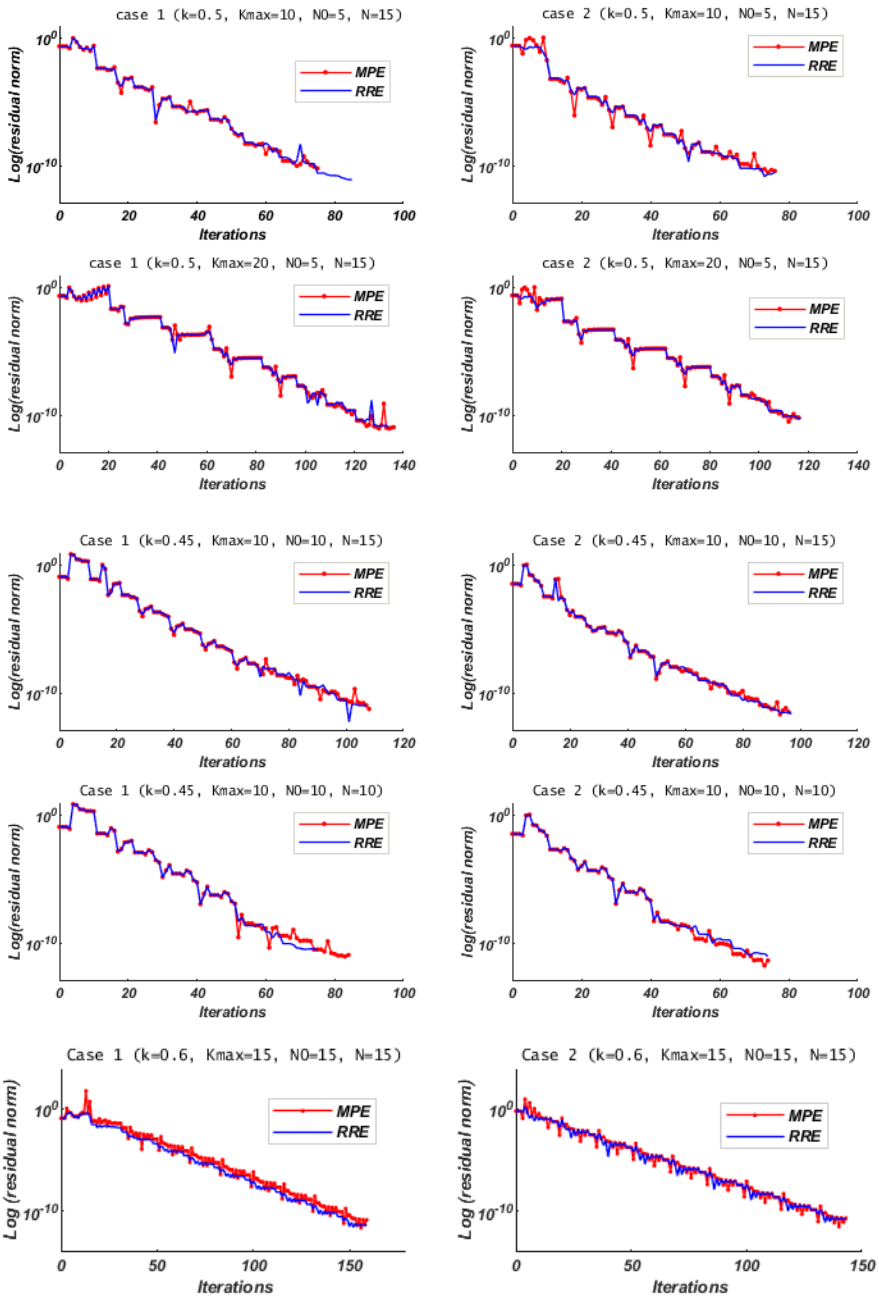

FIGURE 6. Acid Gallic test: Thermodynamic equilibrium by restarted MPE and RRE with $\kappa = 0.45$, $\kappa = 0.5$ and $\kappa = 0.6$ - Residual norm curve

The thermodynamic equilibrium in A is reached after 43 iterations for $m = 1$, 26 iterations for $m = 2$ and 21 iterations for $m \geq 3$ (cf. Figure 7). In B, the convergence of Anderson's method requires 51 iterations for $m = 1$, 30 iterations for $m = 2$, 22 iterations for $m = 3$ and 21 iterations for $m \geq 4$ (cf. Figure 8). In Figures 7 and 8, no complicated oscillation phenomenon is observed, therefore, convergence is achieved without difficulty. Note that concentrations of components at thermodynamic equilibrium in the two mediums A and B are defined respectively by the two vectors:

$$\mathcal{X}_{\mathrm{A}}^{*}\,(\mathrm{M}) = (\mathrm{X}_{1,\mathrm{A}}^{*}, \mathrm{X}_{2,\mathrm{A}}^{*}, \mathrm{X}_{3,\mathrm{A}}^{*}, \mathrm{X}_{4.\mathrm{A}}^{*}, \mathrm{S}_{\mathrm{A}}^{*})^{T} = (10^{-20}, 0.2597, 10^{-20}, 0.3495, 0.3907)^{T}$$
$$\mathcal{X}_{\mathrm{B}}^{*}\,(\mathrm{M}) = (\mathrm{X}_{1,\mathrm{B}}^{*}, \mathrm{X}_{2,\mathrm{B}}^{*}, \mathrm{X}_{3,\mathrm{B}}^{*}, \mathrm{X}_{4.\mathrm{B}}^{*}, \mathrm{S}_{\mathrm{B}}^{*})^{T} = (10^{-20}, 1.5116, 10^{-20}, 0.5756, 7.9128)^{T}$$

(or $\omega_{\mathrm{A}}^{*} = (-20, -0.5855, -20, -0.4565, -0.4081)$ and $\omega_{\mathrm{B}}^{*} = (-20, 0.1794, -20, -0.2399, 0.8983)$ in $\log_{10}$). The influence of the most reactive medium B is demonstrated by the higher concentration of S reached at equilibrium.

For the injection period, we consider two cases, injection in medium A (on the left side) and injection in medium

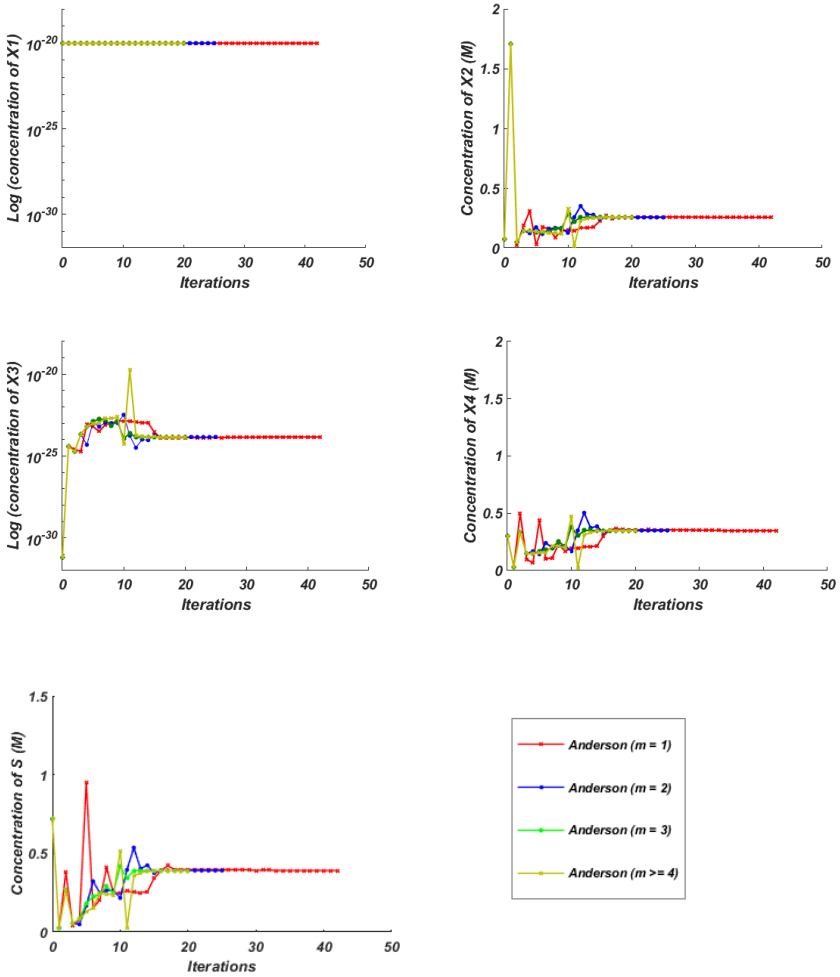

FIGURE 7. MoMas easy test: Thermodynamic equilibrium in medium A by Anderson Acceleration

B. We previously mentioned that the domain is initially in equilibrium with the species S in the presence of mobile species X_2 and X_4. Therefore, for this period, we consider as an initial approximation of the concentrations the thermodynamic equilibrium solution in each medium, namely $\omega_{inj,0} = \begin{cases} \omega_A^* & \text{injection in medium A} \\ \omega_B^* & \text{injection in medium B} \end{cases}$.
Figure 9 and 10 present the behavior of the components concentrations until reaching thermodynamic equilibrium for the injection period in the two zones A and B by Anderson(m) for all $m > 0$. A small number of iterations is necessary to achieve convergence. Note that the equilibrium is reached when the vector of components concentrations is

$$\mathcal{X}_{inj}^*(\mathrm{M}) = (\mathrm{X}_{1,inj}^*, \mathrm{X}_{2,inj}^*, \mathrm{X}_{3,inj}^*, \mathrm{X}_{4,inj}^*, \mathrm{S}_{inj}^*)^T = (0.3, 0.2416, 0.2416, 10^{-50}, 10^{-23})^T$$

(or $\omega_{inj}^* = (-0.5229, -0.6169, -0.6169, -50, -23)^T$ in $\log_{10}$). Once the equilibrium is reached, we notice that the component X_4 is removed to be washed from the domain, X_2 and X_3 remain present in the domain to interact with the surface S and X_4. X_1 is a tracer, which is why its concentration remains constant at 0.3 M.
Leaching follows injection on the same side once the period of 5000 s has passed. To solve the thermodynamic equilibrium for leaching period, an initial approximation the solution of the thermodynamic equilibrium for the injection period is considered, i.e $\mathcal{X}_{leach,0} = \mathcal{X}_{inj}^*$ (or $\omega_{leach,0} = \omega_{inj}^*$).
Figure 11 shows that, with Anderson's method, the thermodynamic equilibrium of the chemical system describing the leaching period is obtained after 61 iterations by Anderson and $m = 2$ and 39 iterations by Anderson and$m = 3$. However, after some numerical tests, we find that for $m = 1$ and $m \geq 4$, Anderson's method does not converge or sometimes presents convergence difficulties. However, after some numerical tests, we find that for $m = 1$ and $m \geq 4$, Anderson's method does not converge or sometimes presents some convergence

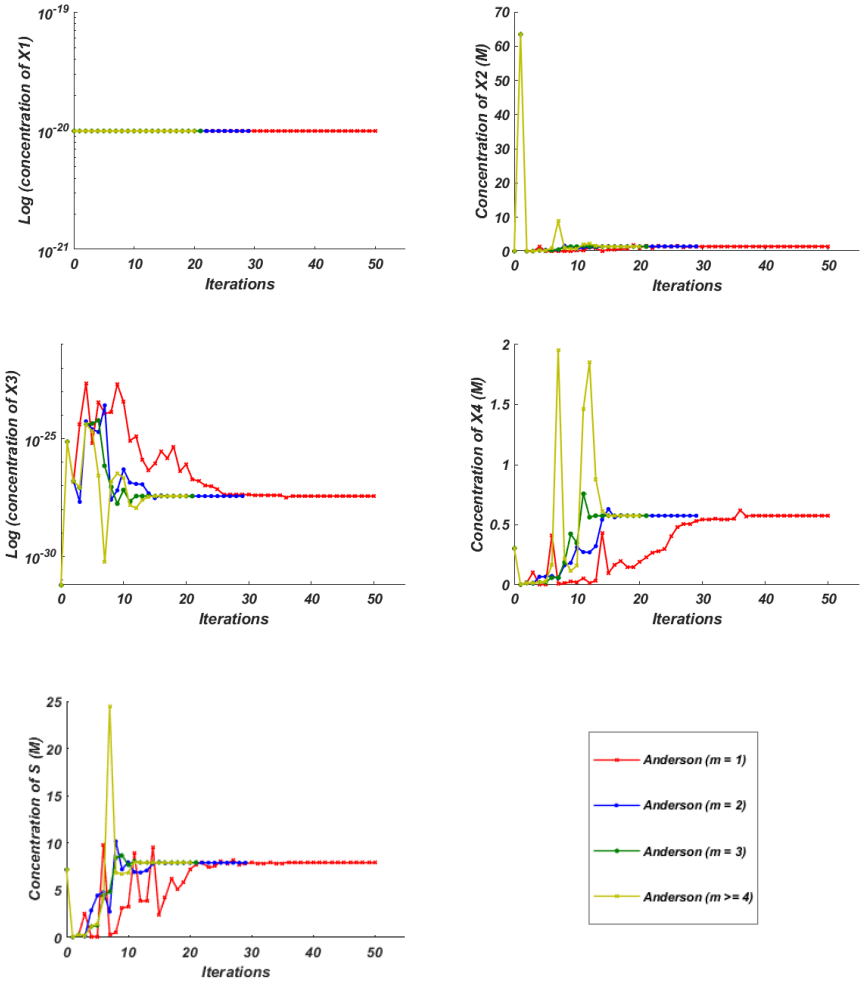

FIGURE 8. MoMas easy test: Thermodynamic equilibrium in medium B by Anderson Acceleration method

difficulties. That comes back, perhaps, to the choice of the initial concentrations. To overcome these difficulties, we consider the fixed point problem (4.1) with relaxation, i.e $\kappa \neq 1$. This parameter is chosen arbitrarily until it is well suited to achieve convergence without difficulty. The convergence of Anderson ($m = 1$), Anderson ($m = 4$), Anderson ($m = 5$) and Anderson($m > 5$) is achieved, in an appropriate number of iterations, taking respectively $\kappa = 0.29$, $\kappa = 0.55$, $\kappa = 0.46$ and $\kappa = 0.42$. The behavior of components concentrations is presented in Figure 11. The solution obtained is

$$\begin{aligned} \mathcal{X}^*_{leach}(\mathrm{M}) &= (\mathrm{X}^*_{1,leach}, \mathrm{X}^*_{2,leach}, \mathrm{X}^*_{3,leach}, \mathrm{X}^*_{4,leach}, \mathrm{S}^*_{leach})^T \\ &= (10^{-20}, 5, 7735.10^{-7}, 7.223.10^{-27}, 1.1547.10^{-6}, 10^{-20})^T \end{aligned}$$

(or $\omega^*_{leach} = (-20, -6.2382, -26.1413, -5.9372, -20)^T$ in $\log_{10}$). We notice that, at the equilibrium of leaching period, 99,9% of the injected pollutant are removed from the domain.
These results again show that the AA method works well for every period and every domain. In addition, the convergence is very fast, requiring a very short computation time (CPU time) not exceeding 2 s. The CPU execution time required by Anderson acceleration method, for several values of the maximal depth m ($m = 1, 2, 3, 4$) to solve the thermodynamic equilibrium in each zone A and B is given in Table 6. This time is given during the two periods of injection and leaching.
On other hand, for the equilibrium in medium A and B, we notice that for $1 \leq m \leq 4$, $\mathbf{cond}(\mathcal{F}_k)$ always remains less than 10^{10}, for any k-th iteration, and it becomes more than 10^{15} for $m \geq 5$ at iteration step 6. With the strategy of condition-number monitoring, $\mathbf{cond}(\mathcal{F}_k)$ returns less than 10^{10} for $k \geq 6$ (cf. Figures 12 and 13). Similarly, by solving the equilibrium system for the injection period in medium A using AA ($m = 1, 2, 3$),

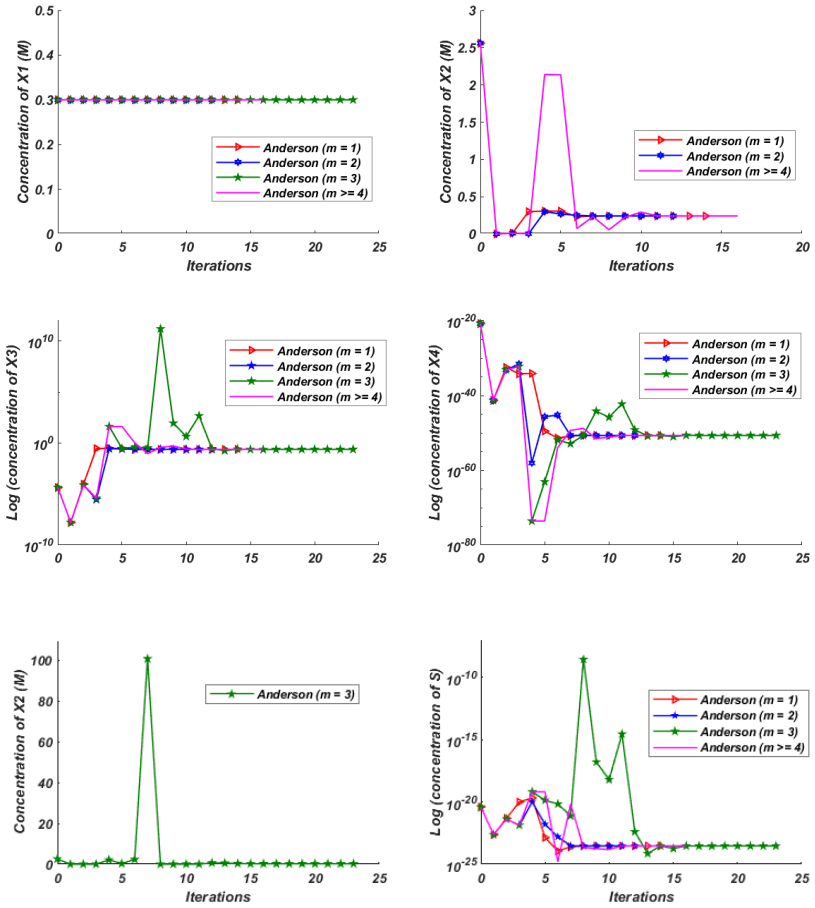

FIGURE 9. MoMas easy test, Injection in medium A: Thermodynamic equilibrium by Anderson Acceleration.

	Anderson (m)	**CPU time (s)**
Zone A	$m = 1$	1.1875
	$m = 2$	1.0156
	$m = 3$	1.0781
	$m = 4$	1.2031
Zone B	$m = 1$	1.6719
	$m = 2$	1.3281
	$m = 3$	1.4063
	$m = 4$	1.6563
Injection period	$m = 1$	1.0156
	$m = 2$	1.0625
	$m = 3$	1.1094
	$m = 4$	1.9844
Leaching period	$m = 1$	1.3594
	$m = 2$	1.4063
	$m = 3$	1.2656
	$m = 4$	1.0313

TABLE 6. MoMas easy test, thermodynamic equilibrium by Anderson Acceleration method: CPU time.

$\mathbf{cond}(\mathcal{F}_k)$ always remains lower than 10^{10}, for any k-th iteration. However, for $m = 4$, at the 13-th iteration, it becomes greater than 10^{10} (cf. Figure 14). By applying the strategy of condition number monitoring to the matrix $\mathcal{F}_{13}$, $\mathbf{cond}(\mathcal{F}_k)$ drops below 10^{10} for $k \geq 13$ and the convergence is reached after 13 iterations.

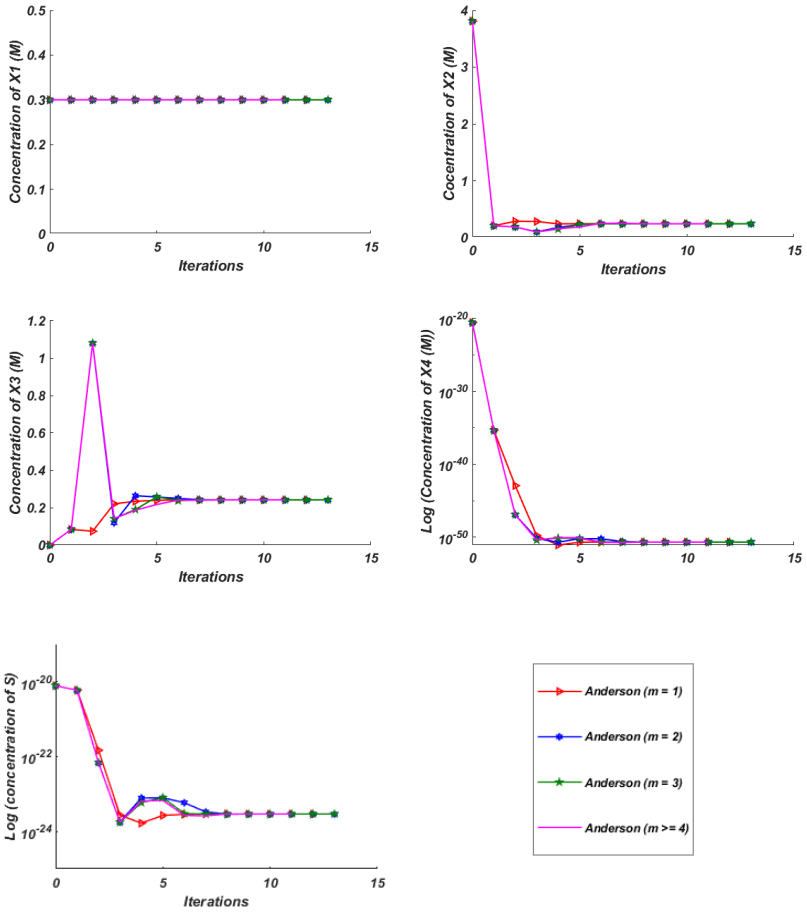

FIGURE 10. MoMas easy test, Injection in medium B: Thermodynamic equilibrium by Anderson Acceleration.

Likewise, for $m \geq 5$, the same strategy is applied to the matrix $\mathcal{F}_6$. In the same way, for the equilibrium chemical systems during the injection period in medium B and the leaching period, the behavior of the condition number of the matrix $\mathcal{F}_k$ with the strategy of condition number monitoring is described in Figures 15 and 16. Note that for the leaching period equilibrium system, the strategy of condition number monitoring is applied at several iteration steps k, to matrices $\mathcal{F}_k$, where $k \in \{11, 12, 13, 15, 17, 18, 22\}$ with Anderson($m = 5$) and $6 \leq k \leq 16, 18 \leq k \leq 20$ with Anderson($m \geq 5$). Therefore, this monitoring strategy has largely contributed to increasing the robustness and stability of the Anderson algorithm.
Figure 17 shows convergence plots with the approximate "theoretical" slopesof Newton for all the cases of MoMas easy test with Anderson method ($m = 1, 2, 3, 4, 5$). For the equilibrium in zones A and B, Anderson($m = 1$) requires twice as many iterations as Anderson($m = 3, 4$). Taking these results into account, the theoretical slopes presented prove the order 2 convergence of the AA method and again demonstrate its efficiency.

We also apply MPE and RRE methods in cyclic mode (through their new implementations described in previous sections) to nonlinear system of thermodynamic equilibrium of MoMas easy test case. The vectors $\omega_1, \omega_2, \ldots$ is generated by (3.11), where the mixing parameter κ is not the same for all the cases of this test. It is chosen arbitrarily to ensure convergence in the most efficient way.
We solve the chemical equilibrium system in medium A by restarted MPE and RRE methods by taking $\kappa = 0.4$. Note that the initial concentrations of component species, for equilibrium system in zones A and B, is given by the vector $\omega_{\text{A,B},0}$ defined above.
The computer program code was run by taking the maximum number of iterations in each cycle $Kmax = 10$. Several choices for the couple (N, N_0): $(N, N_0) = (0, 20)$, $(5, 20)$, $(20, 0)$, $(10, 0)$, $(5, 0)$ are considered. We remind that N_0 is the number of iterations performed before cycling is started, e.g, before MPE or RRE is applied for the first time and N is the number of iterations performed before one of this methods is applied in each cycle after the first cycle.

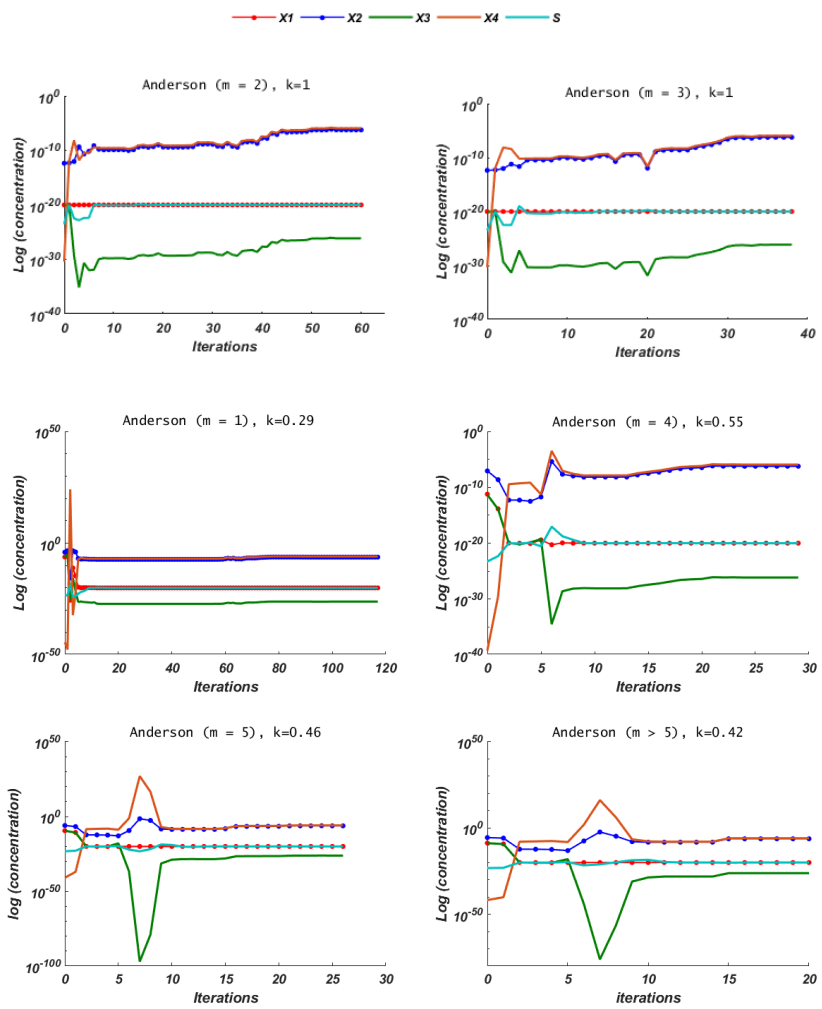

FIGURE 11. MoMas easy test, leaching period: Thermodynamic equilibrium by Anderson Acceleration.

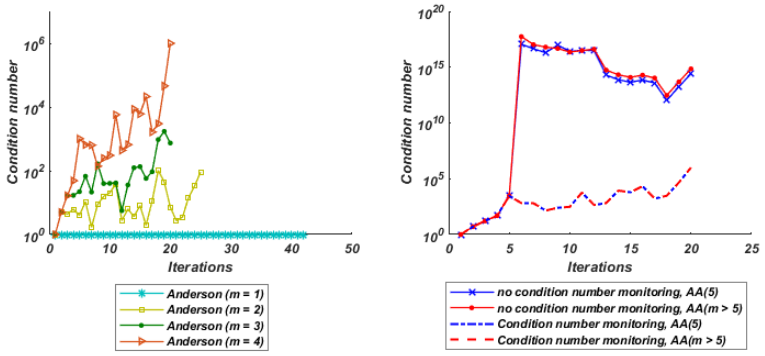

FIGURE 12. MoMas easy test, thermodynamic equilibrium in medium A by Anderson Acceleration method - Condition number curve.

Figure 18 shows the evolution of the nonlinear residual norm, using a logarithmic scale for the restarted MPE and RRE methods. It appears that for $(N, N_0) = (0, 20), (5, 20)$, methods give convergence to a steady state. Moreover, when performing a certain number of iterations before the application of RRE or MPE to each cycle after the first cycle ($N = 5$), the residual norm decreases more rapidly and convergence is reached in a number reduced iterations. For $(N, N_0) = (20, 0)$, MPE converges faster than RRE with few disturbances described by a residual increase between iterations 15 and 20. Finally, we notice that, in the two last cases, RRE and MPE seem to perform similarly and the convergence seems stable.
Regarding the equilibrium in medium B, we propose to take $\kappa = 0.3$. The maximal number of iterations $Kmax$

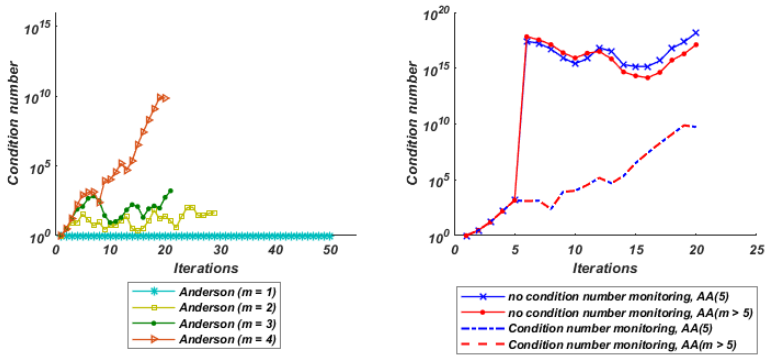

FIGURE 13. MoMas easy test: Thermodynamic equilibrium in medium B by Anderson Acceleration method - Condition number curve.

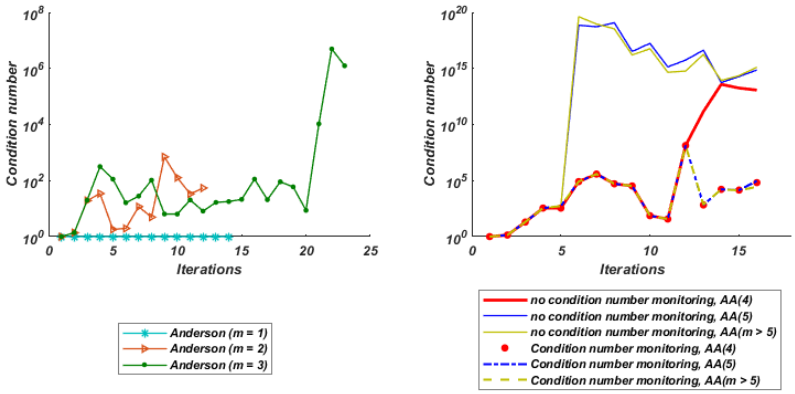

FIGURE 14. MoMas easy test, Injection in medium A: Thermodynamic equilibrium by Anderson Acceleration method - Condition number curve.

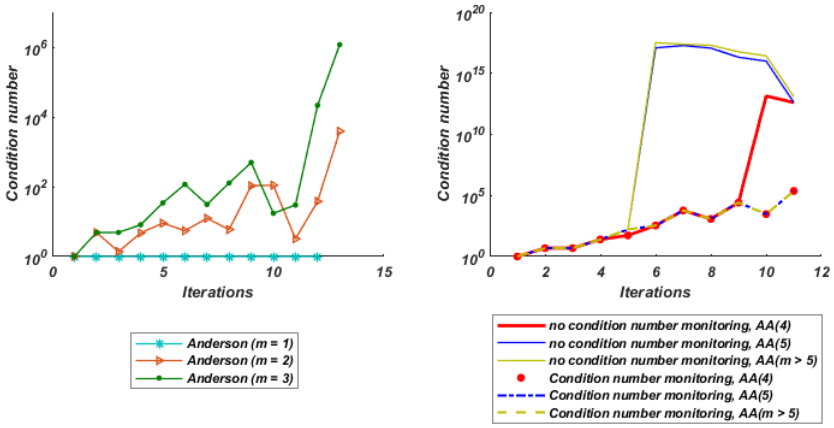

FIGURE 15. MoMas easy test, Injection in medium B: Thermodynamic equilibrium by Anderson Acceleration - Condition number curve.

remains constant equal to 10. Figure 19 contains part of the residual history obtained, for $(N, N_0) = (15, 0)$ and $(N, N_0) = (0, 80)$. It can be seen that, for $(N, N_0) = (0, 80)$, it takes close to 300 iterations to reach the prescribed level of convergence. However, the better case is shown with $(N, N_0) = (15, 0)$ where the RRE and MPE algorithms take approximately 90 and 80 iterations respectively to reach the prescribed levels of residual. Let us emphasize that the numerical experiments carried out for this example show that if one takes $N = 0$, it is necessary to carry out a large number N_0 of iterations before starting the cycling to reach convergence.
Let us now study the thermodynamic equilibrium systems for the injection in media A and B once by taking $\kappa = 0.2$ and once by taking $\kappa = 1$ (i.e without relaxation). Looking now at Figures 20 and 21, it appears that, with the same $Kmax, N_0$ and N, MPE and RRE seem to work very similarly in this example and achieved the same accuracy with both the injection into medium A and with the injection in medium B, except for the choice

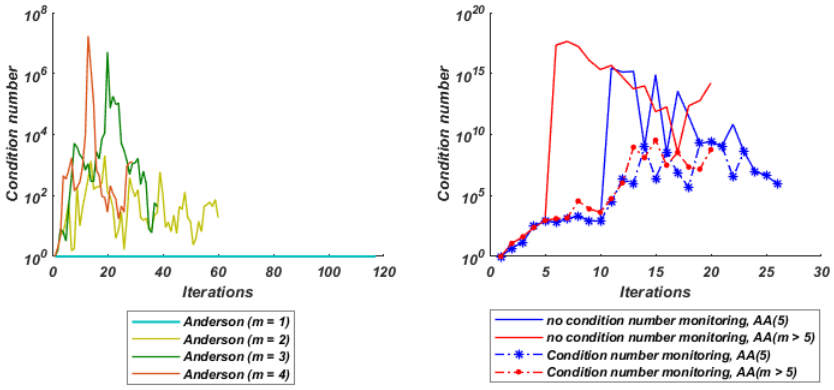

FIGURE 16. MoMas easy test, Leaching period: Thermodynamic equilibrium by Anderson Acceleration method - Condition number curve.

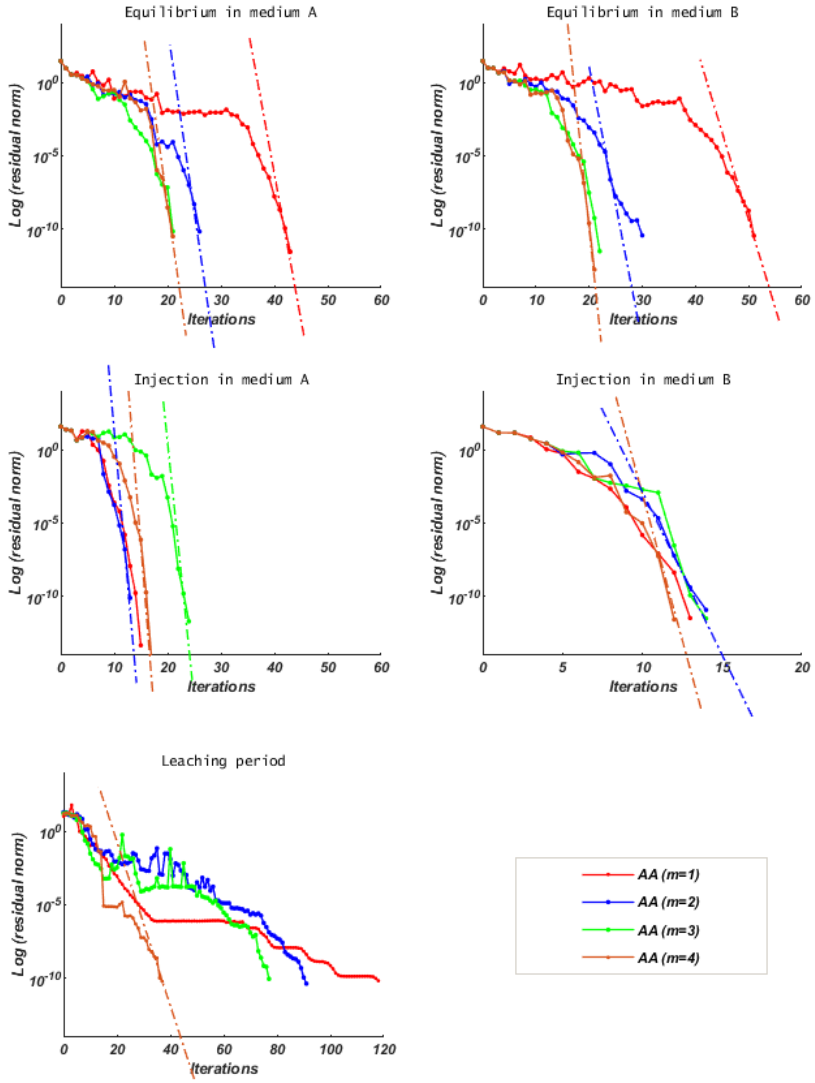

FIGURE 17. MoMas easy test, thermodynamic equilibrium by Anderson Acceleration method -Residual curve.

of $N_0 = 25, N = 0$ in Figure 21. This choice presents a difference between results of two methods: the result of the MPE method decreases compared to the result of RRE, from the fifth cycle and once again resumes its stable and constant trajectory. On the other hand, by taking a relaxation parameter different of 1, we notice that this difference disappears and the two methods again give almost the same results (see Figure 20). Otherwise,

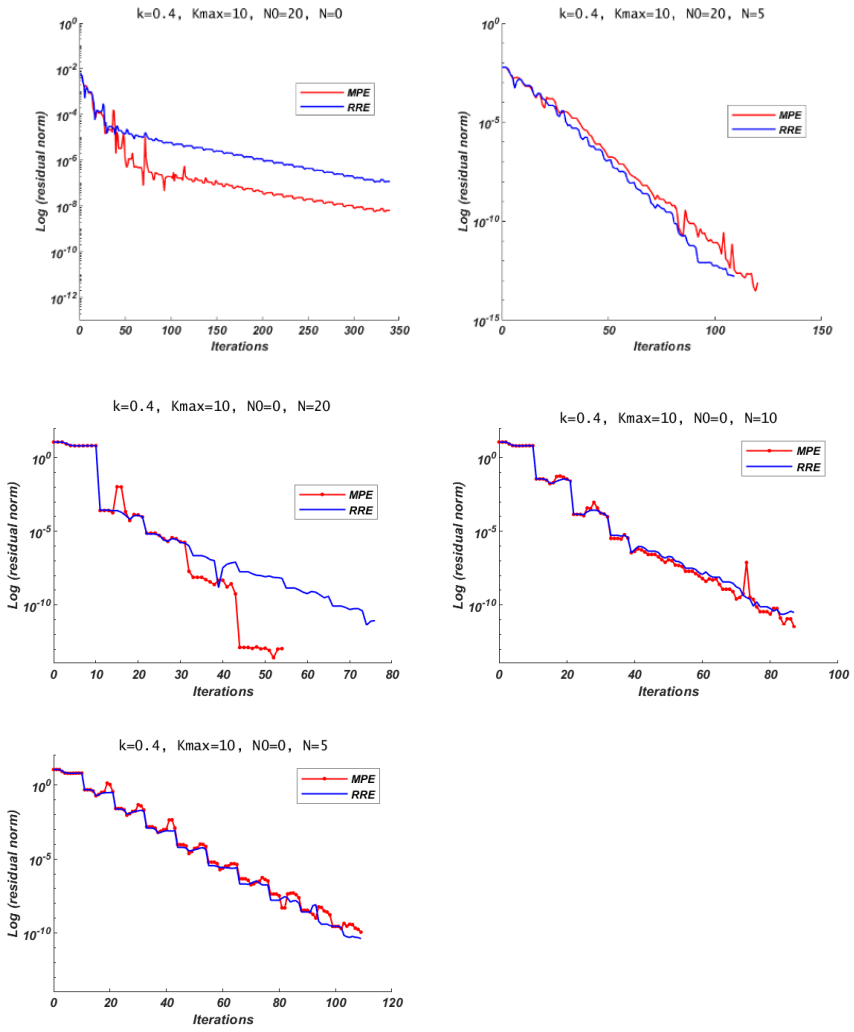

FIGURE 18. MoMas easy test: Thermodynamic equilibrium in the medium A by restarted MPE and RRE, with $\kappa = 0.4$ - Residual norm curve

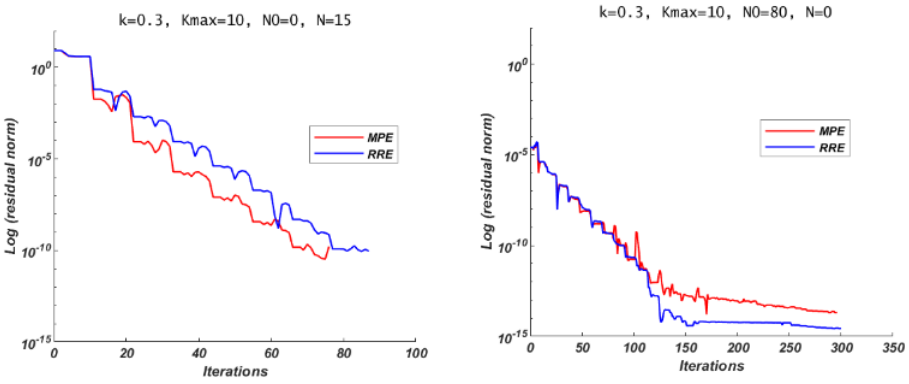

FIGURE 19. MoMas easy test: Thermodynamic equilibrium in the medium B by restarted MPE and RRE, with $\kappa = 0.3$ - Residual norm curve

keeping $\kappa = 1$, this is achieved again by taking the parameter N non equal to zero $(N = 2)$ (cf. Figure22). The parameters involved in the numerical conputation must therefore be chosen with care so that MPE and RRE give consistent results.
Moreover, the number of cycles is reduced when the value of N is increasing, hence taking the time overhead of

cycling into account saves CPU time. This observation confirms the effectiveness concerning the strategy of cycling [44]. Note that for $\kappa = 1$, the convergence is much faster than for $\kappa = 0.2$. A perfect result is obtained for this experiment with $(N, N_0) = (0, 25), (0, 18)$. The prescribed level of convergence for the injection in media A and B respectively is reached very quickly from the first iteration. In addition, with $(N, N_0) = (0, 28), (0, 24)$, the residual norm sometimes seems to be worth a constant lower than the tolerance indicated at this level.

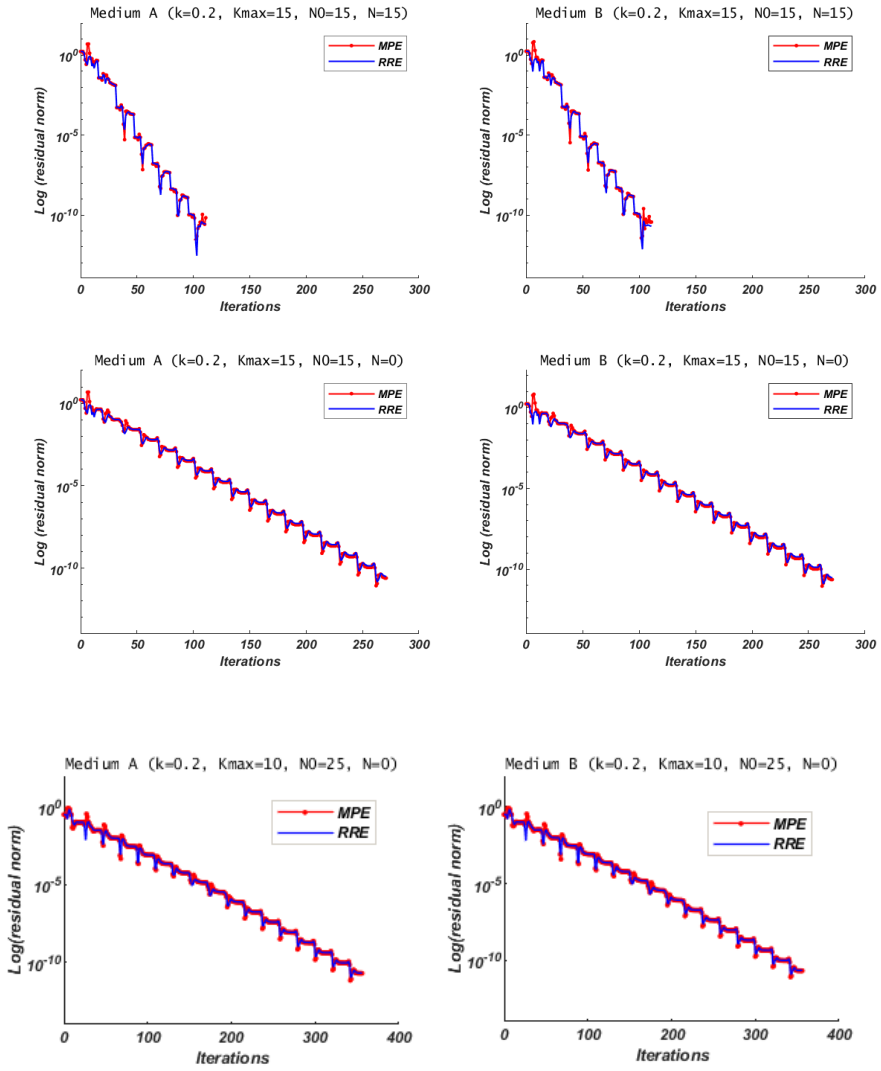

FIGURE 20. MoMas easy test: Thermodynamic equilibrium for injection in media A and B by restarted MPE and RRE, with $\kappa = 0.2$ - Residual norm curve

Finally, it remains to present the resolution of thermodynamic equilibrium by the restarted methods RRE and MPE for the leaching period provided that initial concentrations of components are defined by the solution vector of the equilibrium system for the injection period. We take κ equal to 0.495. This choice for the parameter κ is the best to reach convergence even if it causes difficulties for the convergence. An unstable mode is observed that corresponds to an oscillation of residuals (cf. Figure 23). For $(Kmax, N_0, N) = (10, 0, 10)$, MPE and RRE need the same number of iterations and cycles for convergence, however, for $(Kmax, N_0, N) = (10, 10, 10)$, MPE converges faster than RRE. But, the latter accelerates convergence for $(Kmax, N_0, N) = (10, 5, 18)$ more than MPE. This example appears to be a critical case in that no convergence rule can be deduced by varying the values of N and N_0.

Consequently, to solve the thermodynamic equilibrium system of the leaching period, Anderson Acceleration method appears to be more efficient than the restarted MPE and RRE methods, in particular for a maximal depth $m = 4$. It succeeds in achieving convergence without difficulty, with a stable mode of residual decrease well observed in Figure 17.

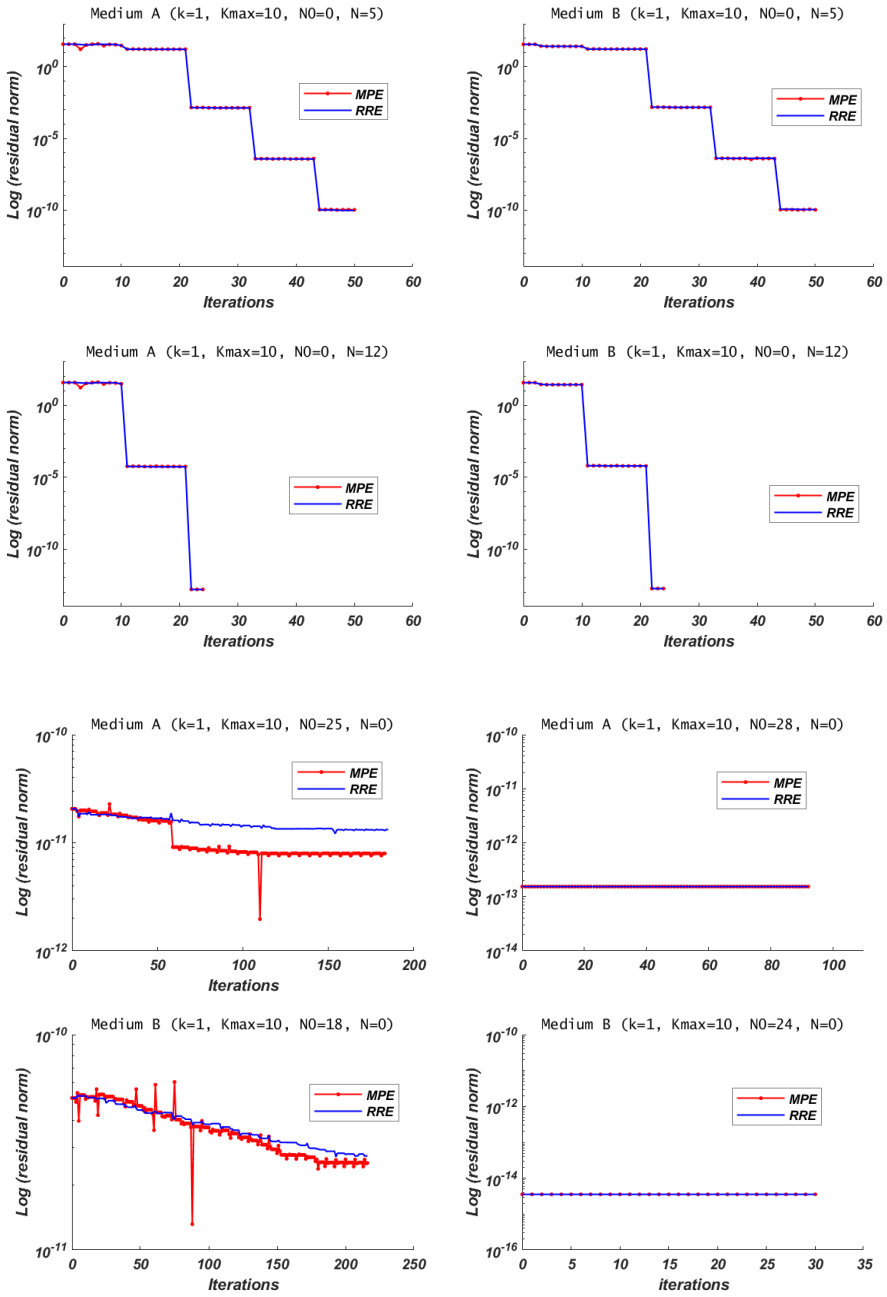

FIGURE 21. MoMas easy test: Thermodynamic equilibrium for injection in media A and B by restarted MPE and RRE, with $\kappa = 1$ - Residual norm curve

All the computation results for the MoMas easy test case are summarized in Table 7. The latter gives the total number of iterations performed $\mathbf{N_{iterations}}$, as well as the number of cycles $\mathbf{N_{cycles}}$ and the computation time CPU necessary for performing the $\mathbf{N_{iterations}}$ iterations and reach convergence. We notice that this time is very short in all cases, not exceeding 3s. This illustrates the efficiency and robustness of the MPE and RRE methods, especially in cyclic mode.

After the computation of components concentrations at equilibrium state, for each period and for each medium, thermodynamic equilibrium concentrations of mobile secondary species (C_1, C_2, C_3, C_4, C_5) can be computed as well as those of fixed secondary species (CS_1, CS_2) thanks to equations in (2.9).

COMPARISON WITH OTHER RESULTS

For the acid Gallic test, the fast convergence observed for the Anderson Acceleration method and for the two polynomial extrapolation methods MPE and RRE is comparable to the results of J. Carrayrou [10]. For the

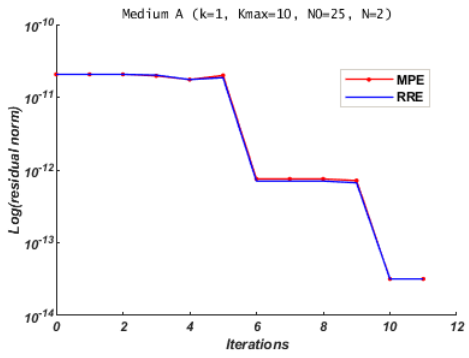

FIGURE 22. MoMas easy test: Thermodynamic equilibrium for injection in medium A by restarted MPE and RRE, $\kappa = 1, N_0 = 25, N = 2, Kmax = 10$ - Residual norm curve

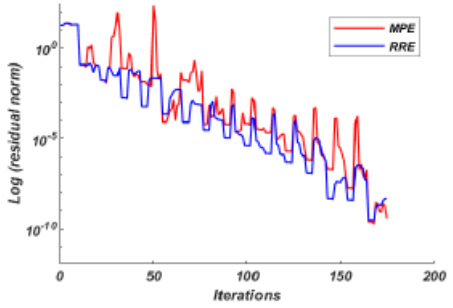

FIGURE 23. MoMas easy test: Thermodynamic equilibrium for leaching period by restarted MPE and RRE - Residual norm curve

	$(\mathbf{Kmax}, \mathbf{N}, \mathbf{N_0})$	$\mathbf{N_{iterations}}$		$\mathbf{N_{cycles}}$		**CPU time (s)**	
		MPE	**RRE**	**MPE**	**RRE**	**MPE**	**RRE**
Zone A, $\kappa = 0.4$	$(10, 0, 20)$	330	330	30	30	1.0469	1.9531
	$(10, 5, 20)$	121	110	11	10	1.3906	0.7969
	$(10, 20, 0)$	55	77	5	7	0.9063	0.7188
	$(10, 10, 0)$	88	88	8	8	0.9844	1.3594
	$(10, 5, 0)$	110	110	10	10	1.0625	0.8750
Zone B, $\kappa = 0.3$	$(10, 15, 0)$	77	88	7	8	1.5469	0.7656
	$(10, 0, 80)$	291	295	30	30	1.7656	1.3438
Injection period, $\kappa = 0.2$	$(15, 15, 15)$	112	112	7	7	1.2188	0.7969
	$(15, 0, 15)$	256	256	17	17	1.5781	1.4688
Injection period, $\kappa = 1$	$(10, 5, 0)$	51	51	5	5	0.8750	0.8594
	$(10, 12, 0)$	25	25	3	3	0.5625	0.7813
	$(10, 0, 25)$ in A	178	180	30	30	1.2031	1.1250
	$(10, 0, 28)$ in A	90	90	30	30	1.2031	0.9844
	$(10, 0, 18)$ in B	210	210	30	30	1.3594	1.3281
	$(10, 0, 24)$ in B	30	30	30	30	0.8594	0.6875
Leaching period, $\kappa = 0.495$	$(10, 10, 0)$	176	176	16	16	1.2813	1.3438
	$(10, 10, 10)$	154	231	14	21	1.4688	2.2656
	$(10, 18, 5)$	154	110	14	10	2.1250	1.5313

TABLE 7. MoMas easy test, thermodynamic equilibrium by restarted MPE and RRE methods.

Newton Raphson type methods, the computation of the thermodynamic equilibrium of the Gallic acid test presents difficulties of convergence. By following the evolution of the process to search a solution in case 1 (cf. Figure 2(a) in [12]), we observe a phenomenon of oscillations during the process of finding solution with Newton Raphson method (which means no convergence). On the other hand, we can note that the Simplex and Newton Raphson with PCF methods allow to obtain an approximation of the solution, without oscillation,

but these require a long computation time. The Simplex method requires significant computation times because the search procedure is far from the solution for a long time. The Newton-Raphson method modified by polishing factor makes it possible to quickly obtain the solution but Figure 2(a) in [12] shows that oscillations are located in a neighborhood close to the solution. It is clear that with Newton-Raphson method modified by imposing the CAI, the oscillations responsible for convergence problems are intrinsic to the CAI procedure. Finally, the Newton-Raphson method with the relaxation by the secant method and the SPECY algorithm allow to effectively approach the solution, avoiding the oscillations and reducing the computation time.
If we compare all the Newton type methods mentioned with the Anderson Acceleration method, the search process is not captured with oscillations. It allows a first accurate approximation of the solution to be obtained more quickly, in a short calculation time (cf. Figure 2 and Table 4). In addition, this method converges in both cases and requires a small number of iterations, for all strictly positive values of the maximal depth m (cf. Figure 2). Likewise, we can quickly get a precise approximation of the solution by applying restarted MPE and RRE methods to the sequence $(\omega_n)_{n\geq 0}$, in both cases, without difficulty and in short computation time.

For the MoMas easy test case, a comparison of our results with those obtained in [27] is presented. The reactive transport code HYTEC participated in the realization of the benchmark, when all chemical reactions are solved by the speciation code CHESS [47]. CHESS uses an improved Newton-Raphson scheme to solve the set of nonlinear algebraic equations describing the chemical system. The HYTEC code was applied to the easy MoMas benchmark as such, without any modification to operate more quickly or to improve convergence, taking the precision of the resolution of chemical equations (Newton Raphson) equal to 10^{-8}.
Far from results concerning transport, we can see in [27] that authors give in a table, results of computation of chemical speciation in initial zones A and B independently, obtained with CHESS code. A comparison be-

	Medium A	Medium B	Injection	Leaching
species				
X_1	1e-20	1e-20	0.3	1e-20
X_2	0.2597	1.5116	0.2416	5.7734e-07
X_3	1.4604e-24	3.6593e-28	0.2416	7.2169e-27
X_4	0.3495	0.5756	2.0800e-51	1.1547e-06
C_1	3.8503e-12	6.6157e-13	4.1387e-12	1.7321e-06
C_2	3.7928e-25	5.5312e-28	0.0584	4.1667e-33
C_3	1.3458	0.3808	8.6087e-51	2
C_4	1.3707e-24	1.3369e-30	6.3800e-152	1e-20
C_5	4.9532e-40	1.4724e-47	1e-20	4.8225e-75
sites				
S	0.3907	7.9128	2.9332e-24	1e-20
CS_2	0.3046	1.0436	1.2687e-97	6e-29
CS_1	9.9968e-21	1e-20	9.9971e-21	1.3889e-59

	zone A	zone B	Injection	Leaching
species				
X_1	-	-	0.3	-
X_2	0.25972	1.5116	0.24162	5.7735e-07
X_3	-	-	0.24162	-
X_4	0.34954	0.57561	-	1.1547e-06
C_1	3.8503e-12	6.6157e-13	4.1387e-12	1.7321e-06
C_2	-	-	0.05838	-
C_3	1.3458	0.38081	-	2
C_4	-	-	-	-
C_5	-	-	-	-
sites				
TS	0.39074	7.9128	-	-
CS_2	0.30463	1.0436	-	-
CS_1	-	-	-	-

TABLE 8. Comparison of the chemical speciation in initial zones obtained by Anderson Acceleration, MPE and RRE methods (on the left) with the chemical speciation obtained by CHESS code (on the right).

tween results obtained with Anderson Acceleration, MPE and RRE with those obtained by the CHESS code (i.e Newton Raphson's method) is summarized in Table 8. One note that the results are the same and in good agreement (all the concentrations lower than 10^{-20} have been represented by "-" in the right part of Table 8).
Four other reactive transport codes also participated in the realization of the benchmark (SPECY, MIN3P, GDAE and *Hoffmann* et *al*), but the results of chemical equilibrium are not presented independently during transport. All these codes are based on a Newton type method to linearize the chemical system and each uses a specific method to find the solution of the linearized system. A reference solution is given by the calculation of SPECY code [11] and a comparison of the results is carried out in [14], where the simulations are given by coupling transport and chemistry. To make a comparison with our results, results of chemistry (our results) on the right and results of the reactive transport on the left are presented. For example, the simulations in [11] indicate that all codes correctly reproduce the increase and decrease of the concentration front C_2. Chemical equilibrium results presented in Figure 24 also exhibit the same behavior for the concentration of C_2 and are in good agreement with those of the reactive transport codes (cf. Fig. 7 in [11]).
On the other hand, chemical equilibrium results obtained for the fixed component S clearly show the influence of the more reactive medium B indicated by the higher concentration of S ($S_B^* = 7.9128$ M). This is in good agreement with the results obtained by the reactive transport codes, in advective case (cf. Fig. 5 in [14]), where at time 10, this high concentration is present in the center of the domain where $1 \leq x \leq 1.1$ (x designates

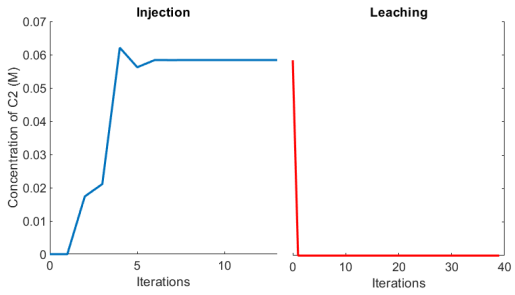

FIGURE 24. MoMas easy test, elution curve for species C2 during injection and leaching periods; chemical equilibrium by Anderson (2), MPE and RRE.

the space), i.e. in B. In addition, we see that in medium A (for $0 \leq x \leq 1$ and $1.1 \leq x \leq 2.1$), the concentration of S converges to the same solution ($S_A^* = 0.39074$ M) that we obtained for chemical equilibrium in A by (AA) (Figure 7) and restarted (MPE) and (RRE) methods.

The most important advantage of the Anderson Acceleration method compared to Newton Raphson type methods is that its algorithm does not require the calculation of the Jacobian matrix. In the resolution of small linear systems using the algorithm of Newton-Raphson, the study of the condition number of Jacobian matrices shows that the range of values covered is unusually large, which leads to specific numerical problems. The matrices are quite small (10×10) but very ill conditioned (up to 10^{200}) (see Table 3 in [28]). This problem is completely overcome with (AA) method when we study the condition-number of the matrix $\mathcal{F}_k$ (or R_k) instead of that of the Jacobian matrix. In addition, the condition-number monitoring strategy used in this method never allows to obtain ill-conditioned matrices. The real $\mathbf{cond}(R_k)$ is always less than 10^{10} with this strategy. In particular, for the Acid Gallic test (respectively MoMas easy test case), with Newton-Raphson type methods, the condition number of the Jacobian matrix varies between $10^{0.61}$ and $10^{12.6}$ (respectively between $10^{3.44}$ and $10^{37.7}$), but with Anderson Acceleration method, the condition-number of matrix $\mathcal{F}_k$ remains less than to 10^{10}, after condition-number monitoring (cf. Figures 3, 12,13, 14, 15 and 16). Then, with (AA) method, efficiency and (relatively) good conditioning can be obtained through updated QR factorizations.
Similarly, for nonlinear problems, the two polynomial-type vector extrapolation methods MPE and RRE do not need the use of the Jacobian of the function $\tilde{\mathbf{G}}$. Moreover, an important property of these methods is that they can be applied directly to the solution of linear and nonlinear systems. This is because the definitions of these methods do not require explicit knowledge of how the sequence is generated.

5. CONCLUSION

The aim of this work is to provide a stable and precise chemical solver to be integrated into an iterative sequential algorithm for reactive transport. The methods presented in this article allow to solve thermodynamic equilibria in a completely new way (without using the Newton-Raphson method). To our knowledge, these iterative acceleration algorithms have never been applied to the resolution of thermodynamic equilibria. The numerical results presented in this article improve the existing results (for example those given in [12]). Thus, the direct combination of the method of positive continued fractions with AA, RRE or MPE provides efficient algorithms with quadratic convergence from any initial arbitrary data.

It is planned to apply these methods to other cases constituting the MoMas reference test cases for which the chemical complexity is increasing (ie "medium test case" and "hard test case"). For this work to bring sufficient novelty, the project consists in coupling our numerical thermodynamic equilibrium resolution to the transport model recently introduced in [4]. In this work, the authors establish a model which describes the water flow in shallow aquifers. The model couples the two dominant flows existing in the aquifer: a vertical 1d-Richards problem is considered in the capillary fringe while a vertical average of the mass conservation law is made in the saturated zone of the aquifer. This study is part of a larger project which aims to model the contamination of groundwater by nitrates.
Obviously, the potential parallelization of the proposed algorithms is an important step in upcoming works, in particular if we want these algorithms to be implemented in the framework of a parallel open-source platform. The parallelization of the MPE and RRE algorithms has already been discussed, in particular in the context of the article [20]. It seems quite possible to adapt these results to our case.

Finally and independently, it would be really very interesting to compare on the problem of thermodynamic equilibria, the results obtained by the AA, RRE or MPE approaches with those obtained thanks to the deep learning methods used in [55, 56].

Acknowledgments: We thank referees as well as Jérôme Carrayrou for their helpful comments and interesting suggestions which allowed to improve the actual version of the article. This work is supported by the project NEEDS-NewSolChem of CNRS.

References

[1] Ahusborde E., Ossmani M. E.., Id Moulay M., *A fully implicit finite volume scheme for single phase flow with reactive transport in porous media*, Mathematics and computers in simulation 164 (2019) 3-23.
[2] Ackerer P., Preface: *Special issue on simulations of reactive transport: Results of the MoMaS benchmarks*, Computat. Geosci. 14(3) (2010), 383.
[3] Anderson, D. G., *Iterative procedures for nonlinear integral equations*. Journal of the ACM, 12 (1965), pp. 547-560.
[4] Bourel, C., Choquet, C., Rosier, C., Tsegmid, M., *Modelling of shallow aquifers in interaction with overland water*. Applied Mathematical Modelling, vol. 81, pp 727-751, (2020)
[5] Bourgeat, A., Bryant, S., Carrayrou, J., Dimier, A., Van Duijn, C.J., Kern, M., Knabner, P., *Benchmark Reactive Transport*. Technical Report GDR MOMAS (2006).
[6] Brassard, P., Bodurtha, P., *A Feasible Set for Chemical Speciation Problems*. Comput. Geosci., 26, 277 (2000).
[7] Brezinski, C., Redivo Zaglia, M., Saad, Y., *Shanks sequence transformations and Anderson acceleration*. SIAM Rev., 60 (2018) 646-669.
[8] Broyden, C. G., *A class of methods for solving nonlinear simultaneous equations*. Math. Comp. 19, 577-593 (1965).
[9] Cabay, S., Jackson, L.W., *A polynomial extrapolation method for finding limits and antilimits of vector sequences*. SIAM J. Numer. Anal. 13 (1976) 734-752.
[10] Carrayrou, J., *Modèlisation du transport de solutés rÃ©actifs en milieu poreux saturé*. Thèse de doctorat, Université Louis Pasteur, Strasbourg, (2001).
[11] Carrayrou, J., *Looking for some reference solutions for the reactive transport benchmark of MoMaS with SPECY*. Computational Geosciences, 14: 393-403, (2010).
[12] Carrayrou, J., Mosé, R., Behra, P., *New efficient algorithm for solving thermodynamic chemistry*. AIChE J. 48(4), 894-904 (2002).
[13] Carrayrou, J., Kern, M., Knabner, P., *Reactive transport benchmark of MoMaS*, Computat. Geosci. 14 (2010), pp. 385-392. 10.1007/s10596-009-9157-7.
[14] Carrayrou, J., Hoffman, J., Knabner, P., Krautle,S. , De Dieuleveult, C., Erhel, J., Van Der Lee, J., Lagneau, V., Mayer, K. U., Macquarrie, K. T. B., *Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems-the MoMaS benchmark case*. Computat. Geosci., 14 (2010), pp. 483-502.
[15] Marinoni M., Carrayrou, J., Lucas Y., Ackerer P., *Thermodynamic equilibrium solutions through a modified Newton Rphson method*, AIChE Journal (2016).
[16] Machat H., Carrayrou, J., *Comparisons of linear solvers for equilibrium geochemistry computatios*, Comput Geosci (2017) 21: 131-150 DOI 10.1007/s10596-016-9600-5
[17] Eddy, R.P., *Extrapolating to the limit of a vector sequence*. in: P.C.C. Wang, Ed., Information Linkage between Applied Mathematics and Industry (Academic Press, New York, 1979) 387-396.
[18] Eyert, V., *A comparative study on methods for convergence acceleration of iterative vector sequences*. J. Comput. Phys. 124(2), 271-285 (1996).
[19] DuMuX, DUNE for Multi-Phase, Component, Scale, Physics, ..., flow and transport in porous media.
[20] Duminil, S., Sadok, H.; Silvester, D. *Fast solvers of discretized Navier-Stokes problems using vector extrapolation*, Numer. Algorithmes 66 (2014), no. 1, 89-104.
[21] DUNE, the Distributed and Unified Numerics Environment, http://www.dune.project.org.
[22] Fang, H., Saad, Y., *Two classes of multisecant methods for nonlinear acceleration*. Numer. Linear Algebra Appl. 16 (3) (2009) 197-221.
[23] Hoffmann J., Kräutle S., Knaber P., *A parallel global-implicit @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 (2010) 421-433.
[24] Jbilou, K., *A general projection algorithm for solving linear systems of equations*. Numer. Algorithms, 4 (1993), pp. 361-377
[25] Jbilou, K., Sadok, H., *Vector extrapolation methods. Application and numerical comparison*. J. Comp. Appl. Math, 122 (2000), 149-165.
[26] Krebs, R., Sardin, M., Schweich, D., *Mineral Dissolution, Precipitation and Ion Exchange in Surfactant Flooding*. AIChE J., 33, 1371 (1987).
[27] Lagneau, V., Van Der Lee, J., *HYTEC results of the MoMas reactive transport benchmark*
[28] Machat, H., Carrayrou, J., *Comparison of linear solvers for equilibrium geochemistry computations*. Comput. Geosci. 21(1) (2017) 131-150.
[29] Mayer K.U., MacQuarrie K.T.B., *Solution of the MoMaS reactive transport benchmark with MIN3P-model formulation and simulation results*, Comput. Geosci. 14 (2010) 405-419.
[30] MeSina, M., *Convergence acceleration for the iterative solution of the equations* $X = AX + f$. Comput. Methocis Appl. Mech. Engrg. 10 (2) (1977) 165-173.
[31] Morin, K. A., *Simplified Explanations and Examples of Computerized Methods for Calculating Chemical Equilibrium in Water*. Comput. Geosci., 11, 409 (1985).
[32] Nelder, J. A., Mead, R., *A Simplex method for function Minimization*. Comput. J., 7, 308 (1965).
[33] 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. Rep. 99-4259, U.S. Geological Survey, Denver, CO (1999).

[34] Potra, F. A., Engler, H., *A characterization of the behavior of the Anderson acceleration on linear problems*. Linear Algebra Appl. 438(3), 1002-1011 (2013).
[35] Pulay, P., *Convergence acceleration of iterative sequences. The case of SCF iteration*. Chem. Phys. Lett. 73(2), 393-398 (1980).
[36] Rohwedder, T., Schneider, R., *An analysis for the DIIS acceleration method used in quantum chemistry calculations*. J. Math. Chem. 49(9), 1889-1914 (2011).
[37] Saad, Y., Schultz, M. H., *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*. SIAM J. Sci. Statist. Comput. 7(3), 856-869 (1986).
[38] Saaltink, MW., Carrera, J., Ayora, C., *Comparison of two approaches for reactive transport modeling*. J. Geochem Explor. (2000); 69:97-101.
[39] Sadok, H.: *About Henrici's transformation for accelerating vector sequences* J. Comput. Appl. Math. 29 (1990) 101-110.
[40] Shapiro, N. Z., Shapley, L. S., *Mass action laws and the gibbs free energy function*. J. SOC. Indust. Appl. Math., 13(2): 353-375, (1965).
[41] Sidi, A., *Convergence and Stability Properties of Minimal Polynomial and Reduced Rank Extrapolation Algorithms*. SIAM J. Numer. Anal. 23, no.1, pp. 197-209, (1986).
[42] Sidi, A., *Extrapolation vs. projection methods for linear systems of equations*. J. Comput. Appl. Math. 22, pp. 71-88, (1988).
[43] Sidi, A., *Efficient implementation of minimal polynomial and reduced rank extrapolation methods*. J. Comput. Appl.Math. vol. 36 (1991), p. 305-337 (cf. p. 32-34).
[44] Sidi, A., *Vector Extrapolation methods with applications to solution of large systems of equations and to PageRank computations*. Computer science department, Technion - Israel institute of technology, Haifa 32000, Israel.
[45] Steefel, C. I., Appelo, C. A. J., Arora, B., Jacques, D., Kalbacher, T., Kolditz, O., Lagneau, V., Lichtner, P. C., Mayer, K., Meeussen, J. C. L., Molins, S., Moulton, D., Shao, H., Simunek, J., Spycher, N., Yabusaki, S. B., Yeh, G. T., *Reactive transport codes for subsurface environmental simulation*. Comput Geosci. (2015); 19(3): 445-478.
[46] Toth, A., Kelley, C. T., *Convergence analysis for Anderson acceleration*. SIAM J. Numer.Anal., 53 (2015), pp. 805-819, https://doi.org/10.1137/130919398.
[47] Van Der Lee, J., *Thermodynamic and mathematical concepts of CHESS*. Technical Report LHM/RD/98/39, CIG- Ecole des l'vlines de Paris, Fontainebleau, France, (1998).
[48] Walker, H.F., *Anderson Acceleration: Algorithms and Implementations*. Research Report, MS-6-15-50, Worcester Polytechnic Institute Mathematical Sciences Department, (2011).
[49] Walker, H.F., Ni, P., *Anderson acceleration for fixed-point iterations*. SIAM J. Numer. Anal. 49 (4) (2011) 1715-1735.
[50] Wood, J. R., *Calculation of Fluid-Mineral Equilibria Using the Simplex Algorithm*. Comput. Geosci., 19, 23 (1993).
[51] Wigley, T. M. L., *WATSPEC: A Computer Program for Determining the Equilibrium Speciation of Aqueous Solutions*. Brit. Geo-morphol. Res. Group Tech. Bull. 20 (1977).
[52] Yeh, G. T., Tripathi, V. S., *A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components*. Water Resources Res. 25, 93-108, (1989).
[53] Yeh, G. T., Tripathi, V. S., Gwo J.P., Cheng H.P., Cheng J.R.C., Salvage K.M., Li M.H., Fang Y., Li Y., Sun J.T., Zhang F. and Siegel M.D., Hydrogeochem: A coupled model of variably saturated flow, thermal transport, and reactive biogeochemical transport. Groundwater Reactive transport models, (2012) 3-41.
[54] Zhang, T., Li, Y., Sun, S., *Phase equilibrium calculations in shale gas reservoirs. Capillarity*, 2019, 2(1): 8-16, doi: 10.26804/capi.2019.01.02.
[55] Zhang, T., Li, Yu, Li Yiteng, Sun, S., Hua B., *A self-adaptive deep learning algorithm for accelerating multi-component flash calculation* , Computer Methods in Applied Mechanics and Engineering (IF 5.763) Volume 369 (2020), DOI: 10.1016/j.cma.2020.113207.
[56] Zhang, T., Li Yiteng, Sun, S., Gao X., *Accelerating flash calculations in unconventional reservoirs considering capillary pressure using an optimized deep learning algorithm*, Journal of Petroleum Science and Engineering, Vol 195 (2020).

[a] Univ. du Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France. [b] CNRS FR 2037, France. [c] Université Libanaise, LaMA-Liban, Laboratoire de recherche en Mathématiques et Applications, P.O. Box 37 Tripoli, Liban.
Email address: Safaa.Al-Nazer@etu.univ-littoral.fr

[c] Université Libanaise, LaMA-Liban, Laboratoire de recherche en Mathématiques et Applications, P.O. Box 37 Tripoli, Liban.
Email address: mjazar@laser-lb.org

[a] Univ. du Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France.[b] CNRS FR 2037, France
Email address: rosier@univ-littoral.fr

Chapitre 3

Modèle couplant le flux de Richards 3D avec le flux horizontal de Dupuit : Cas isotrope et non conservatif

DERIVATION AND MATHEMATICAL ANALYSIS OF DUPUIT-RICHARDS MODEL TAKING INTO ACCOUNT THE FLUID COMPRESSIBILITY.

SAFAA AL NAZER, CAROLE ROSIER, AND MUNKHGEREL TSEGMID

Abstract. In this work, we establish a model which is an alternative to the 3D-Richards equation to describe the flow of water in shallow aquifers. The model couples the two dominant types of flow existing in the aquifer. The first is described by the classic Richards problem in the upper capillary fringe. The second results from Dupuit's approximation after vertical integration of the conservation laws between the bottom of the aquifer and the saturation interface. The final model consists of a strongly coupled system of parabolic-type pde which are defined on a time-dependent domain. First, we show how taking into account the low compressibility of the fluid eliminates degeneration in the time derivative of the Richards equation. Then, we use the general framework of parabolic equations in non-cylindrical domains introduced by Lions to give a global in time existence result to this problem.

Keywords: Richards equation; coupled system of quasilinear parabolic equations; global in time existence; non cylindrical domain.

1. Introduction

Populated areas are increasingly affected by contamination of soil and groundwater. Many modeling approaches are developed to study the vulnerability of aquifers to agricultural pollution, with a particular focus on the supply of nitrates. There is a wide variety of involved processes (chemical, hydrogeological, anthropic, ...) acting in a wide range of temporal and geometrical length scales. But we can notice that the main point for the derivation of the hydrogeological model is linked to a good description of the flow between the ground level (the level of the anthropic processes) and the water table. This will be crucial when studying the transport of chemical components in the aquifer. Indeed, it turns out that many chemical reactions are expected in the first meters of the subsoil, where oxygen is still very present. In particular, chemical species that reach the water table are not necessarily the same as those that have left the surface. This yields different speeds of the reactive kinetics. As a result, for an efficient mathematical modeling, the time upscaling process in this zone must keep track of all the time scales.

In this work, only the hydrogeological question will be considered. Aquifers are often characterized by a form of stratification of flows which enables the definition of interfaces. The slowness of the natural dynamics ensures that the interfaces have a smooth and stable behavior. Besides, due to the dimensions of the aquifer, the flow can be assumed essentially orthogonal to the equipotential (Dupuit's hypothesis). The vertical integration of the Richards equation is thus possible at least in the saturated zone. In this spirit, many 2D models have been developed and used since the 1960s (see for example the works of Jacob Bear, [10, 11]). For more historical notes on the origin of groundwater modeling, we refer interested readers to [16, 17, 20, 27]. But the approach by vertical integration is only valuable for very precise length and time scales, the time scale in particular being completely different of the typical durations of chemical reactions. However, such 2D models are widely used, although it is particularly difficult to correctly couple them to the flow in the unsaturated part of the basement. Several numerical studies have been conducted in this direction. Let us mention the work of [23] where the integrated model is directly coupled with a surface model. In [7], [8] and [18], the coupling of the surface and underground flows is done with a Richards equation associated with a Signorini boundary condition (for the surface behavior). A class of models is proposed in [13] which consists in coupling purely vertical models (for describing the flow at a small time scale) with an horizontal model (describing the flow at a long time scale). They admit the same behavior than the $3D$-Richards model for any time scale when the aquifer present a small deepness compared to its large horizontal dimensions. They describe the essentially horizontal flow of a water table and the essentially vertical water supply flux from the surface through the unsaturated part between the groundwater and the ground level; In [31] we can find a presentation of a rather similar model coupling $1D$-Richards equation with a simplified model in the saturated part. Finally, in [1], this kind of model is integrated into a computational code called "SHE" (for "European Hydrological System" and later became SHETRAN) in

the case where the water table remains away from ground level.

In this paper, we present a model belonging to the "Dupuit-Richards" model class. Indeed, the $3D$-Richards equation is considered in the capillary fringe while a vertical average of the mass conservation law is made in the saturated zone of the aquifer. Pressure and normal fluxes transmission properties are imposed at the saturation interface.
This model differs slightly from that described in [13], already because we consider the complete Richards equations $3D$ in the unsaturated part and no longer only the vertical component of the flow. But the main difference lies in taking into account the low compressibility of the fluid. This will not only make appear a bijective transformation in the time derivative term of the Richards equation but also treat the degeneracy in the parabolic equation governing the horizontal flow in the saturated part of the aquifer. On the other hand, the coupling of flows between the two areas is similar to the one introduced in [13]. It results from the property of continuity of the normal component of the flux at the interface of saturation, ensuring that model is mass conservative.
Of course, the numerical behavior of this model should be similar to that obtained in [13], especially when the horizontal hydraulic conductivity is assumed to be zero in the unsaturated part of the aquifer. Thus, the coupling of the 1D-Richards equation with the 2D-Dupuit approximation numerically justifies this model since from a computation time point of view, it is less expensive than the complete resolution of the 3D-Richards equation.

The mathematical study of the model is particularly delicate because of nonlinearities, the free boundary between each area and the difficulty resulting from the coupling between the two zones which is expressed here by terms of flux at the interface. We must also deal with the mathematical difficulties inherent in Richards equations. Finally, there is a general mathematical difficulty in the structure of the set of PDEs modeling the dynamics of underground water. Indeed, when considering a free water table, we must face the gradual disappearance of water in the desaturation zone and thus the disappearance of a main unknown of the problem.
There exists a huge literature regarding the classical Richards equations. Let us mention the works of Alt *et al* ([4, 5]) and the papers [14, 21, 33] devoted to the study of the degenerate in time equation

$$\partial_t \theta(p) - \Delta p = 0,$$

where $\theta(p)$ denotes the moisture content. We quote also in the one-dimensional case the work of Yin ([38]) concerning the existence of weak solution for the fully degenerate problem

$$\partial_t \theta(p) - \partial_x(\kappa(\theta(p))\partial_x p) = 0,$$

when just assuming that $\theta', \kappa' > 0$.
Classically, the Kirchoff transform is applied to the Richards equation (under appropriate assumptions about porosity and permeability) to eliminate nonlinearity in the diffusive term. In this work, we instead exploit the hypothesis of low compressibility of water to eliminate the degeneracy in the time derivative term of the Richards equation.
This transformation brings us back to the framework of quasilinear parabolic equations on non cylindrical domains to which we can apply the auxiliary domain method introduced by Lions and Mignot [25, 29] to deal with the free boundary.
Besides, vertical averaging in the saturation zone leads to a degenerate elliptic equation whose degeneracy depends on the thickness of the saturated zone. Taking into account the compressibility of the water introduces a degenerescence in the time derivative depending also on the thickness of the saturated zone. A change of variable makes it possible to absorb the two degenerate terms and to return to a regular parabolic equation.

The document is organized as follows: In section 2, we present the first main result of the paper, namely the $3D$-Richards model coupled with the Dupuit horizontal flow; consequences of taking into account the compressibility of the fluid in the modeling are specially detailed. The second main result concerning global in time existence is given in Section 3 as well preliminary results regarding the auxiliary domains method. The proof of the Theorem is performed in section 4. It consists of a fixed point strategy in order to deal with the difficulties linked to nonlinearities and to coupling. The last subsection is devoted to the proof of the existence of an unique solution to the linearized problem with a free boundary by reducing it to a problem in a fixed domain.

2. Derivation of the model

The basis of the modeling is the mass conservation law written for fresh water coupled with the classical Darcy law for porous media. Fluid and soil are considered to be weakly compressible.

For the three-dimensional description, we denote by $\mathbf{x} := (x, z)$, $x = (x_1, x_2) \in \mathbb{R}^2$, $z \in \mathbb{R}$, the usual coordinates.

2.1. Conservation laws

We begin with the conservation of momentum. In view of the (large) dimensions of an aquifer (related to the characteristic size of the porous structure of the underground), we consider a continuous description of the porous medium.

The effective velocity q of the flow is thus related to the pressure P through the Darcy law associated with a nonlinear anisotropic conductivity

$$q = -\frac{\kappa(P)\, K_0}{\mu}(\nabla P + \rho g \nabla z),$$

where ρ and μ are respectively the density and the viscosity of the fluid, K_0 is the permeability of the soil, $\kappa(P)$ is the relative conductivity and g the gravitational acceleration constant. Introducing the hydraulic head H defined by

$$H = \frac{P}{\rho_0 g} + z, \tag{2.1}$$

we write the previous equation as follows:

$$q = -K\nabla H - \frac{\kappa(P)\, K_0}{\mu}(\rho - \rho_0) g \nabla z, \quad K = \frac{\kappa(P)\, K_0 \rho_0 g}{\mu}. \tag{2.2}$$

In this relation, the matrix K is the hydraulic conductivity which expresses the ability of the underground to conduct the fluid. We have denoted by ρ_0 the reference density of the fluid. Next, the conservation of mass during displacement is given by the following equation

$$\partial_t(\theta\rho) + \nabla \cdot (\rho q) = \rho Q, \tag{2.3}$$

where Q denotes a generic source term (for production and replenishment).
The function θ is the volumetric moisture content defined by

$$\theta = \phi s,$$

where ϕ is the porosity of the medium and s is the saturation. If we assume that the air present in the unsaturated zone has infinite mobility, the saturation s and then the function θ are thus considered as monotone functions depending on the pressure as we will detail latter.

2.2. State equation for the fluid compressibility

We consider that the fluid is compressible by assuming that pressure P is related to the density ρ as follows (cf. [15]:

$$\frac{d\rho}{\rho} = \alpha_P dP \Leftrightarrow \rho = \rho_0 e^{\alpha_P (P - P_0)}. \tag{2.4}$$

The real number $\alpha_P \geq 0$ is the fluid compressibility coefficient and P_0 is the pressure of reference. Further assuming $\alpha_P = 0$ we would recover the incompressible case.

2.3. Permeability tensor K_0

The nonlinear hydraulic conductivity K is given by $K = \frac{\kappa(P)\,\rho_0 g}{\mu} K_0$. The soil transmission properties are characterized by the porosity function ϕ and the permeability tensor $K_0(x, z)$. The matrix K_0 is a 3×3 symmetric positive definite tensor which describes the conductivity of the *saturated* soil at the position $(x, z) \in \Omega$. We introduce $K_{xx} \in \mathcal{M}_{22}(\mathbb{R})$, $K_{zz} \in \mathbb{R}^*$ and $K_{xz} \in \mathcal{M}_{21}(\mathbb{R})$ such that

$$K_0 = \begin{pmatrix} K_{xx} & K_{xz} \\ K_{xz}^T & K_{zz} \end{pmatrix}. \tag{2.5}$$

2.4. Hypothesis

Let us now list the assumptions on the fluid and medium characteristics but also on the flow which are meaningful in the context of our problem.

Hypothesis on the fluid and on the medium

Soil Compressibility We neglect in the model the effects of the rock compressibility, the porosity of the medium ϕ do not depend on the pressure variations and it is thus assumed to be a constant.

Compressibility of the fluid First, we assume that the fluid (namely here fresh water) is weakly compressible. It means that

$$\alpha_P \ll 1. \tag{2.6}$$

Let us exploit this assumption. In natural conditions and especially in an aquifer, one observes small fluid mobility (defined by the ratio κ/μ). First consequence of the low compressibility of the fluid combined with the low mobility of fluid appears in the momentum equation. We perform a Taylor expansion with regard to P of the density ρ in the gravity term of the Darcy equation. Neglecting the terms weighted by $\alpha_P\kappa/\mu \ll 1$ in (2.2), we get:

$$q = -K\nabla H, \quad K = \frac{\kappa(P)\,\rho_0 g}{\mu} K_0. \tag{2.7}$$

Second consequence is $\nabla\rho \cdot q \ll 1$ which leads to the following simplification in the mass conservation equation (2.3):

$$\rho\partial_t\theta + \theta\,\partial_t\rho + \rho\nabla\cdot q = \rho Q.$$

Neglecting in this way the variation of density in the direction of flow is sometimes considered as an extra assumption called Bear's hypothesis (*cf* [2]). Including (2.4), that is $\partial_t\rho = \rho\alpha_P\partial_t P$ in the latter equation, we get

$$\rho\partial_t\theta + \rho\,\theta\alpha_P\,\partial_t P + \rho\nabla\cdot q = \rho Q.$$

After simplification by $\rho > 0$, we finally obtain

$$\partial_t\theta + \theta\alpha_P\,\partial_t P + \nabla\cdot q = Q. \tag{2.8}$$

Equivalently, using the hydraulic head (2.1) and the Darcy law (2.7), (2.8) can be written

$$\partial_t\theta \;+\; S_0\partial_t H - \nabla\cdot(K\nabla H) = Q \quad \text{where} \quad S_0 = \rho_0\, g\,\phi\,\alpha_P. \tag{2.9}$$

We notice that if the fluid is assumed incompressible, $\alpha_P = 0$, then Eq. (2.8) is the classical Richards equation in pressure formulation. An adequate definition of the volumetric moisture content θ and of the mobility function κ is the key of the model.

Richards hypothesis. The Richards model is moreover based on the assumption that the air pressure in the underground equals the atmospheric pressure, thus is not an unknown of the problem. One thus assumes that the saturation and the relative conductivity of the soil are given as *functions* of the fluid pressure P, denoted respectively by $s = s(P)$ and $\kappa = \kappa(P)$. We introduce the saturation pressure P_s which is a fixed real number. The fully-saturated part of the medium corresponds to the region $\{\mathbf{x}; P(.,\mathbf{x}) > P_s\}$, while it is partially-saturated in the capillary fringe $\{\mathbf{x}; P_d < P(.,\mathbf{x}) \le P_s\}$. The dry part is defined by the set $\{\mathbf{x}; P(.,\mathbf{x}) \le P_d\}$. The moisture content is such that

$$\theta = \begin{cases} \phi & \text{(saturated zone)} & \text{if} \quad P(.,\mathbf{x}) > P_s, \\ \theta(P) & \text{(with } 0 \le \theta(P) \le \phi \text{ and } \theta'(P) > 0\text{)} & \text{if} \quad P_d < P(.,\mathbf{x}) \le P_s, \\ \theta_0 = \phi s_0 & \text{(dry zone)} & \text{if} \quad P(.,\mathbf{x}) \le P_d, \end{cases} \tag{2.10}$$

where $s_0 > 0$ corresponds to a residual saturation which is positive. The associated relative hydraulic mobility is then defined by

$$\kappa(P) = \begin{cases} 1 & \text{(saturated zone)} & \text{if} \quad P(.,\mathbf{x}) > P_s, \\ \kappa(\theta(P)) & \text{(with } 0 \le \kappa(P) \le 1 \text{ and } (\kappa o\theta)'(P) > 0\text{)} & \text{if} \quad P_d < P(.,\mathbf{x}) \le P_s, \\ 0 & \text{(dry zone)} & \text{if} \quad P(.,\mathbf{x}) \le P_d. \end{cases} \tag{2.11}$$

There is a large choice of available models for s and κ. The most classical examples for an air-water system are the van Genuchten model [37] with no-explicit dependance on the bubbling pressure but with fitting parameters, and the Brooks and Corey model [9].

The important point is that these models are such that

$$s(P) = 1 \iff P \ge P_s \quad \text{and} \quad \kappa(P) = 1 \iff P \ge P_s. \tag{2.12}$$

In particular, the water pressure is greater than the bubbling pressure P_s if and only if the soil is completely saturated .

HYPOTHESIS ON THE FLOW

The following assumption is introduced for upscaling the 3D problem to a 2D model in the saturated part of the domain.

Dupuit approximation (hydrostatic approach) Dupuit assumption consists in considering that the hydraulic head is constant along each vertical direction (vertical equipotentials). It is legitimate since one actually observes quasi-horizontal displacements when the thickness of the aquifer is small compared to its width and its length and when the flow is far from sinks and wells.

2.5. GEOMETRY

The aquifer is represented by a three-dimensional domain $\Omega := \Omega_x \times (h_{bot}, h_{soil})$, $\Omega_x \subset \mathbb{R}^n$ with $n \geq 2$ ($x = (x_1, x_2)$), function h_{bot} (respect. h_{soil}) describing its lower (respect. upper) topography. The upper and lower surfaces are thus defined by the graph of the functions $h_{bot} = h_{bot}(x)$ and $h_{soil} = h_{soil}(x)$, $x \in \Omega_x$. We assume that

$$h_{\rm soil}(x) > h_{\rm bot}(x), \qquad \forall x \in \Omega_x. \tag{2.13}$$

More precisely the domain is given by:

$$\Omega = \Big\{(x,z) \in \Omega_x \times \mathbb{R} \quad | \quad z \in \big]h_{\rm bot}(x), h_{\rm soil}(x)\big[\Big\}. \tag{2.14}$$

We always denote by $\vec{\nu}$ the outward unit normal and $\vec{e_3}$ is the unitary vertical vector pointing up. We decompose the boundary $\partial\Omega$ of Ω in three zones (bottom, top and vertical)

$$\partial\Omega = \Gamma_{\rm bot} \sqcup \Gamma_{\rm soil} \sqcup \Gamma_{\rm ver}\ ,$$

with

$$\Gamma_{\rm bot} := \Big\{(x,z) \in \Omega \,|\, z = h_{\rm bot}(x)\Big\}\,, \quad \Gamma_{\rm soil} := \Big\{(x,z) \in \Omega \,|\, z = h_{\rm soil}(x)\Big\},$$
$$\Gamma_{\rm ver} := \Big\{(x,z) \in \Omega \,|\, x \in \partial\Omega_x\Big\}.$$

Our model split the description of the flow in two subregions of Ω (possibly time-dependent) in each of which the flow presents different behavior. We denote by h the depth of the free interface separating the freshwater layer and the unsaturated part of the aquifer. The definition of these zones is thus based on the function $h = h(t,x)$ which is an unknown of our problem. We then introduce, for a given function $h = h(t,x)$ such that $h_{\rm bot} \leq h \leq h_{\rm soil}$:

$$\Omega_t^- := \big\{(x,z) \in \Omega \,|\, z < h(t,x)\big\} \quad \text{and} \quad \Omega_t := \big\{(x,z) \in \Omega \,|\, z > h(t,x)\big\}, \tag{2.15}$$

and

$$\Gamma_t := \big\{(x,z) \in \Omega \,|\, z = h(t,x)\big\}. \tag{2.16}$$

2.6. MODEL COUPLING VERTICAL 3D-RICHARDS FLOW AND DUPUIT HORIZONTAL FLOW

- **Three-dimensional Richards equation in the upper capillary fringe**

In the unsaturated part of the aquifer, Ω_t, the 3D-Richards equation (2.8) holds

$$\begin{cases} \partial_t\theta + \theta\alpha_P\,\partial_t P + \nabla\cdot q = Q & \text{for } (t,x,z) \in (0,T)\times\Omega_t, \\ q\cdot\vec{\nu} = 0 & \text{for } (t,x,z) \in (0,T)\times\big(\Gamma_{\rm soil}\cup\Gamma_{\rm ver}\big), \\ P\big(t,x,h(t,x)\big) = P_s & \text{for } (t,x) \in (0,T)\times\Omega_x, \\ P(0,x,z) = P_{\rm init}(x,z) & \text{for } (x,z) \in \Omega_0. \end{cases} \tag{2.17}$$

The effective velocity q is given by

$$q = -K\nabla\Big(\frac{P}{\rho_o\, g} + z\Big), \quad K = \frac{\kappa(P)\,K_0\rho_0 g}{\mu}.$$

We emphasize that the model (2.17) depends by definition on the depth h which is expected to belong to the interval (h_{bot}, h_{soil}).

- **Dupuit horizontal flow in the saturated zone**

Upscaling procedure

We now use the approximations introduced in (2.4) to vertically integrate equation (2.9), thus reducing the 3D problem to a 2D problem. We perform the vertical integration between depths h_{bot} and h. Since $\theta(P) = \phi$ in the saturated zone, the vertical average (2.9) leads to

$$\int_{h_{bot}}^{h} \left(S_0 \partial_t H + \nabla \cdot q\right) dz = \int_{h_{bot}}^{h} Q \, dz.$$

We denote by $B_f = h - h_{bot}$ the thickness of the saturated zone and by $\tilde{Q}$ the source term representing distributed surface supply of fresh water into the free aquifer:

$$\tilde{Q} = \frac{1}{B_f} \int_{h_{bot}}^{h} Q \, dz.$$

Applying Leibnitz rule to the first term in the left-hand side yields:

$$\int_{h_{bot}}^{h} S_0 \partial_t H dz = S_0 \frac{\partial}{\partial t} \int_{h_{bot}}^{h} H dz \, - S_0 H_{|z=h} \partial_t h + S_0 H_{|z=h_{bot}} \partial_t h_{bot}.$$

We denote by $\tilde{H}$ the vertically averaged hydraulic head

$$\tilde{H} = \frac{1}{B_f} \int_{h_{bot}}^{h} H dz.$$

Because of Dupuit approximation, $H(x_1, x_2, z) \simeq \tilde{H}(x_1, x_2)$, $x = (x_1, x_2) \in \Omega$, $z \in (h_{bot}, h)$, we have

$$\int_{h_{bot}}^{h} S_0 \partial_t H dz = S_0 B_f \partial_t \tilde{H}.$$

We also have

$$\int_{h_{bot}}^{h} \nabla \cdot q \, dz = \nabla' \cdot (B_f \tilde{q'}) + q_{|z=h^-} \cdot \nabla(z-h) - q_{|z=h_{bot}^+} \cdot \nabla(z - h_{bot}),$$

where $\nabla' = (\partial_{x_1}, \partial_{x_2})$, $q' = (q_{x_1}, q_{x_2})$ and the averaged Darcy velocity $\tilde{q'} = \frac{1}{B_f} \int_{h_{bot}}^{h} q' \, dz$ is given by

$$\tilde{q'} = -\frac{1}{B_f} \int_{h_{bot}}^{h} \left(K \nabla' H\right) dz = -\frac{1}{B_f} \int_{h_{bot}}^{h} \left(K \nabla' \tilde{H}\right) dz = -\tilde{K} \nabla' \tilde{H}, \quad \tilde{K} = \frac{1}{B_f} \int_{h_{bot}}^{h} \frac{K_0 \rho_0 g}{\mu} \, dz,$$

(we remind that $\kappa(P) = 1$ for $z \in (h_{bot}, h)$). The averaged mass conservation law for the freshwater in the saturated zone finally reads

$$S_0 B_f \partial_t \tilde{H} = \nabla' \cdot (B_f \tilde{K} \nabla' \tilde{H}) + q_{|z=h_{bot}^+} \cdot \nabla(z - h_{bot}) - q_{|z=h^-} \cdot \nabla(z-h) + B_f \tilde{Q}. \tag{2.18}$$

In this equation, term $B_f \tilde{K}$ may be viewed as the dynamic transmissivity of freshwater layer. At this point, we have obtained an undetermined system of two pdes ((2.17)-(2.18)) with three unknowns P, $\tilde{H}$ and h.

Fluxes and continuity equations across the interface

Our aim is now to include in the model the continuity and transfert properties across interface. As a consequence, we express the two flux terms appearing in (2.18) and we reduce the number of unknowns.

- Flux across the saturation interface:
  The saturation interface is characterized by the cartesian equation $F(x_1, x_2, z, t) = 0 \Leftrightarrow z - h(x_1, x_2, t) = 0$, the unit normal vector $\vec{\nu}$ to the interface is thus colinear to $\nabla(z-h)$.
  The relation ruling continuity of the normal component of the velocity thus reads

$$\left(q_{|z=h^+} - q_{|z=h^-}\right) \cdot \vec{\nu} = 0 \Leftrightarrow q_{|z=h^+} \cdot \nabla(z-h) = q_{|z=h^-} \cdot \nabla(z-h). \tag{2.19}$$

- Approximation of the flux $q_{|z=h^+} \cdot \nabla(z-h)$:
  The flux $q_{|z=h^+} \cdot \nabla(z-h)$ expresses mass transfers between the two parts of the aquifer. As it is done in [13], we approximate the flux by

$$q_{|z=h^+} \cdot \nabla(z-h) \simeq \int_{h(t,x)}^{h_{soil}(x)} \left(\phi \frac{\partial s(P)}{\partial t} + \phi \, s(P) \alpha_P \frac{\partial P}{\partial t} - Q\right) dz. \tag{2.20}$$

This approximation comes from the hypothesis of an almost null horizontal hydraulic conductivity (i.e. $K_{xx} \simeq (0)$) in the capillary fringe. It corresponds to a flow almost vertical in this part of the aquifer. So the 3D-Richards equation is reduced to a 1D-equation. Integrating this $1D$ equation between h and h_{soil} yields the approximation (2.20).
It is an essential difference with the mathematical analysis presented in [36] in which the exchanges between the two parts of the aquifer were simplified and represented by the addition of an external source term, thus decoupling the two problems.

- Impermeable layer at $z = h_{soil}$
  Since the lower layer is impermeable, there is no flux across the boundary $z = h_{bot}$:

$$q(h_{bot}) \cdot \nabla(z - h_{bot}) = 0. \tag{2.21}$$

- Continuity equations:
  Continuity relation imposed on the interface enables to properly reduce the number of unknowns in equations (2.17)-(2.18).
  Dupuit approximation reads $\tilde{H} \simeq H_{|z=h^-}$, the pressure P thus satisfies in Ω_t^-

$$P(t,x,z) = \rho_0\, g\big(\tilde{H}(t,x) - z\big) \qquad \text{for } t \in [0,T[\,, \quad (x,z) \in \Omega_t^-. \tag{2.22}$$

  Besides, the pressure is continuous across Γ_t, it follows that

$$P(t,x,h^-) = P(t,x,h^+) = P_s \Leftrightarrow \tilde{H} = \frac{P_s}{\rho_0 g} + h. \tag{2.23}$$

Equation (2.23) allows to substitute $\tilde{H}$ by h in Eq. (2.18), we thus have

$$S_0 B_f \partial_t h - \nabla' \cdot (B_f \tilde{K} \nabla' h) = B_f \tilde{Q} \tag{2.24}$$
$$- \int_{h(t,x)}^{h_{soil}(x)} \big(\phi\, \frac{\partial s(P)}{\partial t} + \phi\, s(P) \alpha_P \frac{\partial P}{\partial t} - Q\big) dz \ in\ (0,T) \times \Omega_x,$$
$$\tilde{K} \nabla' h \cdot \vec{\nu} = 0 \text{ on } (0,T) \times \partial\Omega_x, \tag{2.25}$$

with

$$B_f = (h - h_{bot}), \quad \tilde{K} = \frac{1}{B_f} \int_{h_{bot}}^{h} \frac{K_0 \rho_0 g}{\mu}\, dz \quad \text{and} \quad S_0 = \rho_0\, g\, \phi\, \alpha_P. \tag{2.26}$$

The homogeneous Neumann condition on $\partial\Omega_x$ is assumed to simplify the presentation.
The final model $(\mathcal{M})$ coupling 3D-Richards flow and Dupuit horizontal flow consists in system (2.17), (2.22) and (2.24), namely we have

- In Ω_t the following 3d-Richards equation holds

$$\begin{cases} \partial_t \theta(P) + \theta \alpha_P\, \partial_t P + \nabla \cdot q = Q & \text{in} \quad (0,T) \times \Omega_t, \\ q \cdot \vec{\nu} = 0 & \text{on} \quad (0,T) \times \big(\Gamma_{\text{soil}} \cup \Gamma_{\text{ver}}\big), \\ P\big(t,x,h(t,x)\big) = P_s & \text{in} \quad (0,T) \times \Omega_x, \\ P(0,x,z) = P_0(x,z) & \text{in} \quad \Omega_0. \end{cases}$$

  The effective velocity q is given by

$$q = -K \nabla\Big(\frac{P}{\rho_o\, g} + z\Big), \quad K = \frac{\kappa(P)\, K_0 \rho_0 g}{\mu}.$$

- In Ω_t^- the pressure P satisfies

$$P(t,x,z) = \rho_0\, g\big(\frac{P_s}{\rho_0 g} + h - z\big) \quad \text{in} \quad (0,T) \times \Omega_t^-.$$

- The depth of Γ_t, h, satisfies in Ω_x

$$\begin{cases} S_0 B_f \partial_t h - \nabla' \cdot (B_f \tilde{K} \nabla' h) = & B_f \tilde{Q} - \int_{h(t,x)}^{h_{soil}(x)} \big(\phi\, \frac{\partial s(P)}{\partial t} + \phi\, s(P) \alpha_P \frac{\partial P}{\partial t} - Q\big)\, dz, \\ \tilde{K} \nabla' h \cdot \vec{\nu} = 0 & \text{on} \quad (0,T) \times \partial\Omega_x, \\ h(0,x) = h_0(x) & \text{in} \quad \Omega_x. \end{cases}$$

3. Mathematical setting and main results

The problem (2.17) being a problem with free boundary, we are going to define the general framework of parabolic equation in non cylindrical domain, introduced by Lions and Mignot respectively in [25] and [29].

3.1. Notations and auxiliary results

Let $\mathcal{O}$ be the open domain of $\mathbb{R}\times\mathbb{R}^N$, included in the set $\mathbb{R}^+\times\Omega$ defined by

$$\mathcal{O}=\mathbb{R}^+\times\Omega_x\times(h,h_{soil}),$$

where h is the position of the interface Γ_t. We set

$$\begin{aligned}
&\Omega_{t'}=\mathcal{O}\cap\{t=t'\},\ \forall t'\geq 0,\ \ \text{(definition compatible with (2.15))}\\
&\Omega'_{t'}=\Omega\setminus\Omega_{t'},\ \forall t'\geq 0,\\
&\mathcal{O}_T=\mathcal{O}\cap\{0\leq t\leq T\},\\
&\mathcal{O}'_T=\big((0,T)\times\Omega\big)\setminus\mathcal{O}_T,\\
&\Gamma'=\Gamma\setminus\Omega_0\ \ \text{(i.e. the lateral boundary of }\mathcal{O}),\\
&\gamma_t=\partial\Omega_t\ \ \text{(i.e. the boundary of }\Omega_t\subset\mathbb{R}^N),\\
&\Gamma'_T=\Gamma'\cap\{0<t<T\}=\cup_{t\in(0,T)}\gamma_t.
\end{aligned}$$

We define

$$H^{0,1}(\mathcal{O})=\{u\,|\,D^pu\in L^2(\mathcal{O})\ \text{for}\ |p\,|\leq 1\},$$

where

$$D^pu=\{D^\alpha u|\alpha=(\alpha_1,\alpha_2,\alpha_2)\text{with}\ |\alpha|=p\}.$$

It is an Hilbert space endowed with the norm

$$\|u\|_{H^{0,1}(\mathcal{O})}=\big(\Sigma_{p\leq 1}\int_{\mathcal{O}}|D^pu|dxdt\big)^{1/2}.$$

$H_0^{0,1}(\mathcal{O})$ denotes the closure of $\mathcal{D}(\mathcal{O})$ in $H^{0,1}(\mathcal{O})$ for the norm $\|\ \|_{H^{0,1}(\mathcal{O})}$. In the following, we will denote $F(\mathcal{O})=H_0^{0,1}(\mathcal{O})$ and $F'(\mathcal{O})$ its topological dual. Besides, we introduce

$$\begin{aligned}
\mathcal{A}(\mathcal{O})&=\{u|\,u\in H^{0,1}(\mathcal{O}),\ \frac{du}{dt}\in F'(\mathcal{O})\},\\
\mathcal{B}(\mathcal{O})&=\{u|\,u\in F(\mathcal{O}),\ \frac{du}{dt}\in F'(\mathcal{O})\},
\end{aligned}$$

endowed with the Hilbertian norms
$\|\cdot\|_{\mathcal{A}(\mathcal{O})}=\big(\|\cdot\|^2_{H^{0,1}(\mathcal{O})}+\|\partial_t\cdot\|^2_{(H^{0,1}(\mathcal{O}))'}\big)^{1/2}$ and $\|\cdot\|_{\mathcal{B}(\mathcal{O})}=\big(\|\cdot\|^2_{H_0^{0,1}(\mathcal{O})}+\|\partial_t\cdot\|^2_{F'(\mathcal{O})}\big)^{1/2}$.
Finally, $B_0(\mathcal{O})$ (resp. $B_T(\mathcal{O})$) is the closure in $\mathcal{B}(\mathcal{O})$ of functions null in a neighborhood of $t=0$ (resp. t=T).
We now state some auxiliary results proved in [25]

Lemma 3.1. *If $\mathcal{O}$ is sufficiently regular, we thus have*

- *1. $H^{0,1}(\mathcal{O}_T)=L^2([0,T];H^1(\Omega_t))$ where*

  $$L^2([0,T];H^1(\Omega_t))=\{u\,|u(t,.)\in H^1(\Omega_t),\ t\in[0,T]\ \ a.e.\ and\ \|u\|_{H^{0,1}(\mathcal{O}_T)}<+\infty\},$$

  with $\|u\|_{H^{0,1}(\mathcal{O}_T)}=\int_0^T\|u\|^2_{H^1(\Omega_t)}\,dt$.
  A similar result holds for $H^{0,1}(\mathcal{O}_T)$.
- *2. For $u\in H_0^{0,1}(\mathcal{O})$, we can define $\gamma(u)$, the trace of u on Γ' in $L^2(\Gamma')$.*
  Moreover $u\in F(\mathcal{O})\Longleftrightarrow\gamma(u)=0$.
- *3. Let $u\in\mathcal{B}(\mathcal{O}_T)$, thus $u\in B_T(\mathcal{O})\Longleftrightarrow u(T,.)=0$.*
- *4. $\forall u,v\in\mathcal{B}(\mathcal{O}_s)$, we have*

  $$\langle\frac{\partial u}{\partial t},v\rangle_{F',F}+\langle u,\frac{\partial v}{\partial t}\rangle_{F',F}=(u(s,.),v(s,.))_{L^2(\Omega_s)}-(u(0,.),v(0,.))_{L^2(\Omega_0)}.\tag{3.1}$$

Let Ω' an open bounded domain of $\mathbb{R}^3$. For the sake of brevity we shall write $H^1(\Omega')=W^{1,2}(\Omega')$ and

$$V(\Omega')=H_0^1(\Omega'),\ V'(\Omega')=H^{-1}(\Omega'),\ H(\Omega')=L^2(\Omega').$$

The embeddings $V(\Omega') \subset H(\Omega') = H'(\Omega') \subset V'(\Omega')$ are dense and compact. For any $T > 0$, let $W(0,T,\Omega')$ denote the space

$$W(0,T,\Omega') := \{\omega \in L^2(0,T;V(\Omega')),\ \partial_t\omega \in L^2(0,T;V'(\Omega'))\}$$

endowed with the Hilbertian norm $\|\cdot\|_{W(0,T,\Omega')} = \left(\|\cdot\|^2_{L^2(0,T;V(\Omega'))} + \|\partial_t \cdot\|^2_{L^2(0,T;V'(\Omega'))}\right)^{1/2}$. The following embeddings are continuous ([26] prop. 2.1 and thm 3.1, chapter 1)

$$W(0,T,\Omega') \subset \mathcal{C}([0,T];[V(\Omega'),V'(\Omega')]_{\frac{1}{2}}) = \mathcal{C}([0,T];H(\Omega'))$$

while the embedding

$$W(0,T,\Omega') \subset L^2(0,T;H(\Omega')) \tag{3.2}$$

is compact (Aubin's Lemma, see [35]).
The following result by F. Mignot (see [22]) is used in the sequel.

Lemma 3.2. *Let $f : R \to R$ be a continuous and nondecreasing function such that $\limsup_{|\lambda|\to+\infty} |f(\lambda)/\lambda| < +\infty$.*
Let $\omega \in L^2(0,T;H(\Omega'))$ be such that $\partial_t\omega \in L^2(0,T;V'(\Omega'))$ and $f(\omega) \in L^2(0,T;V(\Omega'))$.
Then

$$\langle \partial_t\omega, f(\omega)\rangle_{V(\Omega')',V(\Omega')} = \frac{d}{dt}\int_\Omega \left(\int_0^{\omega(\cdot,y)} f(r)\,dr\right) dy \text{ in } \mathcal{D}'(0,T).$$

Hence for all $0 \le t_1 < t_2 \le T$

$$\int_{t_1}^{t_2} \langle \partial_t\omega, f(\omega)\rangle_{V',V}\,dt = \int_\Omega \left(\int_{\omega(t_1,y)}^{\omega(t_2,y)} f(r)\,dr\right) dy.$$

Remark 3.3. The result (3.1) of Lemma 3.1 is a generalization of Lemma 3.2 to the case where the space domain Ω' is time dependent.

3.2. Main results

We aim giving an existence result of physically admissible weak solutions for model $(\mathcal{M})$ completed by initial and boundary conditions.
Let us first detail the mathematical assumptions. We begin with the characteristics of the porous structure. We limit our study to the isotropic case so K_0 is assumed to be a scalar. In the saturated part, the averaged hydraulic conductivity $\tilde{K}$ is thus equal to the constant $\frac{K_0\rho_0 g}{\mu}$. Without loss of generality, we will assume a zero source term in Eq. (2.24) (that is $\tilde{Q} = 0$). Functions θ and κ are pressure-dependent and we assume

$$\theta \in \mathcal{C}^1(\mathbb{R}), \quad 0 < \theta_- := \phi s_0 \le \theta(x) \le \theta_+, \quad \theta'(x) \ge 0\ \forall x \in \mathbb{R}, \tag{3.3}$$

$$\kappa \in \mathcal{C}(\mathbb{R}), \quad 0 < \kappa_- \le \kappa(x) \le \kappa_+\ \forall x \in \mathbb{R}. \tag{3.4}$$

Before stating the main result of this work, we will transform the original problem and bring us back to the framework introduced in [29].
The above assumptions on the fluid and the medium allow to eliminate the nonlinearity in time of Eq. (2.17), namely Assumptions (3.3)-(3.4) are sufficient to define the primitive function $\mathcal{P}$ such that

$$\mathcal{P}(P) = \theta(P) + \alpha_P \int^P \theta(s)\,ds.$$

A direct computation gives $\mathcal{P}'(P) = \theta'(P) + \alpha_P\theta(P) > \alpha_P\theta_- > 0$, indeed by previous hypothesis, we have $\theta'(P) \ge 0$ and $\theta(P) > \phi s_0$.
Since $\mathcal{P}$ is a bijective application, the existence of p such that

$$p = \mathcal{P}(P)$$

is equivalent to the existence of P solution of the original Richards problem. The transform $\mathcal{P}$ of Eq. (2.17) is

$$\partial_t p - \frac{1}{\mu}\nabla\cdot\left(\frac{\kappa(\mathcal{P}^{-1}(p))}{(\theta' + \alpha_P\theta)(\mathcal{P}^{-1}(p))}K_0\nabla p\right) - \frac{\rho_0\, g}{\mu}\nabla\cdot\left(\kappa(\mathcal{P}^{-1}(p))K_0\vec{e_3}\right) = Q.$$

Finally, we introduce the notation

$$\tau(p) = \frac{K_0}{\mu}\frac{\kappa(\mathcal{P}^{-1}(p))}{(\theta' + \alpha_P\theta)(\mathcal{P}^{-1}(p))}.$$

Note that, due to hypotheses (3.3)-(3.4), there exist two positive reals τ_- and τ_+ such that

$$0 < \tau_- := \frac{K_0\,\kappa_-}{\mu\,\alpha_P\theta_+} \leq \tau(p) \leq \tau_+ := \frac{K_0\,\kappa_+}{\mu\,\alpha_P\theta_-}. \tag{3.5}$$

Let $\delta \in \mathbb{R}$ be a positive number and $l = (h_{soil} - h_{bot})$ the function (space depending) denoting to the total thickness of the subsoil. We introduce the function T_l defined by

$$T_l(u) = \sqrt{u} + h_{bot} \quad \forall u \in [\delta^2, l^2],$$

which is extended continuously and constantly outside $[\delta^2, l^2]$.

Remark 3.4. In order to extend the solution p outside the time dependent domain Ω_t, it is necessary to impose on the function h to be less than or equal to a quantity strictly greater than h_{bot}. This is the reason why the small parameter δ was introduced.

Let $\chi_0(u)$ the function defined by

$$\chi_0(u) = \begin{cases} 0 & \text{if} \quad u \leq 0 \\ 1 & \text{if} \quad u > 0 \end{cases}.$$

In order to guarantee the non-negativity of u, the control $\chi_0(u)$ is added in front of the right hand side of (3.6) . Setting $u = (h - h_{bot})^2$, Eq. (2.24) thus becomes

$$\frac{S_0}{2}\partial_t u - \frac{\tilde{K}}{2}\nabla' \cdot (\nabla' u) = -\chi_0(u) \int_{T_l(u(t,x))}^{h_{soil}(x)} \left(\frac{\partial p}{\partial t} - Q\right) dz. \tag{3.6}$$

Definition 3.5. *The definition of the depth h is derived from the construction of u. Namely, for u given by (3.6), we set*

$$h(t,x) := T_l(u). \tag{3.7}$$

Remark 3.6. This definition of h allows to define the integration domain Ω_t (and then the interface Γ_t) in the system (3.8)-(3.9). We emphasize that by definition, h always remains in the interval $[h_{\rm bot} + \delta, h_{\rm soil}]$.

We are led to consider the following problem completed by the boundary and initial conditions :

$$\partial_t p - \nabla \cdot \left(\tau(p)\,\nabla p\right) - \nabla \cdot \left(\kappa(\mathcal{P}^{-1}(p))K_0\vec{e_3}\right) = Q \quad \text{in} \quad \mathcal{O}_T, \tag{3.8}$$

$$p\big|_{\Gamma_t} = \mathcal{P}(P_s) \quad \text{in} \quad (0,T), \qquad \nabla\left(\mathcal{P}^{-1}(p) + \rho_0 g z\right) \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times (\Gamma_{soil} \cup \Gamma_{ver}),$$

$$p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0. \tag{3.9}$$

$$\frac{S_0}{2}\partial_t u - \frac{\tilde{K}}{2}\nabla' \cdot (\nabla' u) = -\chi_0(u) \int_{T_l(u(t,x))}^{h_{soil}(x)} \left(\frac{\partial p}{\partial t} - Q\right) dz \quad in \qquad (0,T) \times \Omega_x, \tag{3.10}$$

$$\nabla u \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times \partial\Omega_x, \quad u(0,x) = (h_0(x) - h_{bot}(x))^2 \quad \text{in} \quad \Omega_x. \tag{3.11}$$

where P_s is constant with respect to the time and the space. Function $P_0 \in H^2(\Omega)$ satisfies the compatibility condition

$$P_0(x, h_0) = P_s \quad \text{in} \quad \Omega_0.$$

We also assume that $h_0 \in L^\infty(\Omega_x)$ is such that

$$h_{bot} + \delta \leq h_0 \leq h_{soil} \text{ a.e. in} \quad \Omega_x. \tag{3.12}$$

Source term Q is given function of $L^2(0,T;H(\Omega))$. For the previous parabolic system, we state and prove the following existence result.

Theorem 3.7. *Assume that there exist two real numbers θ_- and κ_- such that*

$$\theta(x) \geq \theta_- > 0 \quad \forall x \in \mathbb{R}, \quad \kappa(x) \geq \kappa_- > 0 \quad \forall x \in \mathbb{R}^+. \tag{3.13}$$

Then system (3.8)- (3.9),(3.10)-(3.11) admits a weak solution (p,u) satisfying
(a) the function $p \in L^2(0,T;H^1(\Omega)) \cap L^2(0,T;(H^1(\Omega))')$ is solution of (3.8)- (3.9);
(b) the function $u \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ is solution of (3.10)- (3.11). Moreover $u(t,x) \geq 0$ a.e. in $[0,T] \times \Omega_x$.

Corollary 3.8. *Assume that there exist two real numbers θ_- and κ_- such that*

$$\theta(x) \geq \theta_- > 0 \quad \forall x \in \mathbb{R}, \quad \kappa(x) \geq \kappa_- > 0 \quad \forall x \in \mathbb{R}^+. \tag{3.14}$$

Then the model $\mathcal{M}$ admits a weak solution (P, h) such that
(a) the function $P \in L^2(0,T; H^1(\Omega)) \cap L^2(0,T; (H^1(\Omega))')$;
(b) the function $h \in L^2(0,T; H^1(\Omega_x)) \cap L^2(0,T; (H^1(\Omega_x))')$ and $h(t,x) \in [h_{\rm bot} + \delta, h_{\rm soil}]$ a.e. in $[0,T] \times \Omega_x$.

The proof of Corollary 3.8 is a direct consequence of Theorem 3.7 since we turn back to the original problem by considering the inverse transform $\mathcal{P}^{-1}$.

Next section is devoted to the proof of Theorem 3.7.

4. Proof of Theorem 3.7

Let us sketch the global strategy of the proof. The problem consists of a strongly nonlinear coupled system, so we apply a fixed point approach to solve it in two steps. In the first step, we decouple the system and apply a fixed point Schauder theorem to establish an existence and uniqueness result for each decoupled and regularized equation. Then we establish compactness results which allow us to prove the global existence in time of the initial problem. One of the main difficulties of the study is that we are working on time-dependent domains. This difficulty is solved by using the work of Lions and Mignot for parabolic equations on non-cylindrical domains. This consists in suitably extending the solution outside the variable domain, thus bringing us back to a fixed domain ([25, 29]).

4.1. Fixed point step

We first reduce the boundary condition on interface Γ_t of the system (3.8)-(3.9) to homogeneous Dirichlet boundary condition. To do this, we set $\bar{p} = p - \mathcal{P}(P_s)$. Since $\mathcal{P}(P_s)$ is a constant, the system (3.8)-(3.9) becomes

$$\begin{aligned}
&\partial_t \bar{p} - \nabla \cdot \big(\bar{\tau}(\bar{p})\, \nabla \bar{p}\big) - \frac{\rho_0\, g}{\mu} \nabla \cdot \Big(\bar{\kappa}(\bar{p}) K_0 \vec{e_3}\Big) = Q \quad \text{in} \quad \mathcal{O}_T,\\
&\bar{p}\big|_{\Gamma_t} = 0 \quad \text{in} \quad (0,T), \qquad \nabla\big(\mathcal{P}^{-1}\big(\bar{p} + \mathcal{P}(P_s)\big) + \rho_0 g z\big) \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times (\Gamma_{soil} \cup \Gamma_{ver}),\\
&\bar{p}(0,x,z) = \mathcal{P}(P_0)(x,z) - \mathcal{P}(P_s) \quad \text{in} \quad \Omega_0,
\end{aligned}$$

where $\bar{\tau}(\bar{p}) = \tau\big(\bar{p} + \mathcal{P}(P_s)\big)$ and $\bar{\kappa}(\bar{p}) = \kappa o \mathcal{P}^{-1}\big(\bar{p} + \mathcal{P}(P_s)\big)$. We thus remark, that just renaming functions τ and κ, we go back to the case $\mathcal{P}(P_s) = 0$ on Γ_t. So, from now, we omit the subscript "$\bar{\cdot}$" in the previous system and we consider the original system (3.8)-(3.9) with $\mathcal{P}(P_s) = 0$.

$$\partial_t p - \nabla \cdot \big(\tau(p)\, \nabla p\big) - \frac{\rho_0\, g}{\mu} \nabla \cdot \Big(\kappa(\mathcal{P}^{-1}(p)) K_0 \vec{e_3}\Big) = Q \quad \text{in} \quad \mathcal{O}_T, \tag{4.1}$$

$$p\big|_{\Gamma_t} = 0 \quad \text{in} \quad (0,T), \qquad \nabla\big(\mathcal{P}^{-1}(p) + \rho_0 g z\big) \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times (\Gamma_{soil} \cup \Gamma_{ver}),$$

$$p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0. \tag{4.2}$$

Definition 4.1. *We call weak solution of problem (4.1)-(4.4) any solution $p \in W(0,T,\Omega)$ s.t.*
1) $p = 0$ in $\Omega \setminus \Omega_t$, $\forall t \in (0,T)$,
2) the solution p satisfies the weak formulation in $\mathcal{O}_T$, $\forall \phi \in \mathcal{A}(\mathcal{O})$ (null on the interface Γ_t)

$$\langle \partial_t p, \phi \rangle_{F',F} + \int_0^T \Big(\int_{\Omega_t} \big(\tau(p) \nabla p + \frac{\rho_0\, g}{\mu} \kappa(\mathcal{P}^{-1}(p)) K_0 \vec{e_3}\big) \cdot \nabla \phi \Big) - Q\phi \Big)\, dt = 0, \tag{4.3}$$

$$p(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \textit{in} \quad \Omega_0. \tag{4.4}$$

We now construct the framework to apply the Schauder fixed point theorem (see [19, 39]). For the fixed point strategy, we introduce two convex subsets $(W_1,\, W_2)$ of $W(0,T,\Omega_x) \times W(0,T,\Omega)$, namely

$$W_1 \;:=\; \{u \in W(0,T,\Omega_x);\; u(0) = u_0, \|u\|_{L^2(0,T;H^1(\Omega_x))} \leq C_u \text{ and } \|u\|_{L^2(0,T;(H^1(\Omega_x))')} \leq C_u'\}.$$

and

$$W_2 \;:=\; \{p \in W(0,T,\Omega);\; p(0) = p_0, \|p\|_{L^2(0,T;H^1(\Omega))} \leq C_p \quad \text{and} \quad \|p\|_{L^2(0,T;(H^1(\Omega))')} \leq C_p'\},$$

constants $(C_p,\ C'_p)$ and $(C_u,\ C'_u)$ being defined thereafter.
Let $(\bar u, \bar p) \in W_1 \times W_2$, we begin by considering the unique solution u of the following linearized problem

$$\frac{S_0}{2}\partial_t u - \frac{\tilde K}{2}\nabla'\cdot(\nabla' u) = -\chi_0(u)\int_{\bar h(t,x)}^{h_{soil}(x)} \big(\frac{\partial \bar p}{\partial t} - Q\big)\,dz \quad in \qquad (0,T)\times\Omega_x, \tag{4.5}$$

$$\nabla u\cdot\vec\nu = 0 \quad \text{on} \quad (0,T)\times\partial\Omega_x, \quad u(0,x) = (h_0(x) - h_{bot}(x))^2 \quad \text{in} \quad \Omega_x, \tag{4.6}$$

where $\bar h(t,x) := T_l(\bar u(t,x))$.

Remark 4.2. 1) We have to precise the meaning of the term $\int_{\bar h(t,x)}^{h_{soil}(x)} \frac{\partial \bar p}{\partial t}\,dz$:

$$\int_{\bar h(t,x)}^{h_{soil}(x)} \frac{\partial \bar p}{\partial t}\,dz = \int_{h_{bot}(x)}^{h_{soil}(x)} \chi_{z\geq \bar h(t,x)}\frac{\partial \bar p}{\partial t}\,dz$$

is the function of $(H^1(\Omega_x))'$ such that $\forall v \in H^1(\Omega_x) \subset H^1(\Omega)$

$$\langle \int_{\bar h(t,x)}^{h_{soil}} \frac{\partial \bar p}{\partial t} dz,\ v\rangle_{H^1(\Omega_x)',H^1(\Omega_x)} = \langle \frac{\partial \bar p}{\partial t}, \chi_{z\geq \bar h(t,x)}\ v\rangle_{H^1(\Omega)',H^1(\Omega)}.$$

2) Note that thanks to the change of variable $u = (h - h_{bot})^2$, Eq. (2.24) (which is nonlinear and degenerate in space and time) has now a parabolic structure.

Lemma 4.3. *Let $h_0 \in L^\infty(\Omega_x)$ satisfying (3.12), there exists a unique weak solution $u \in W(0,T,\Omega_x)$ of (4.5)-(4.6) such that*

$$\|u\|_{L^2(0,T;H^1(\Omega_x))} \leq C_u \quad \textit{and} \quad \|u\|_{L^2(0,T;(H^1(\Omega_x))')} \leq C'_u,$$

where C_u and C'_u only depend on the data of the problem.
Moreover, $u \geq 0$, a.e. in $[0,T]\times\Omega_x$.

Proof. STEP 1. GLOBAL EXISTENCE AND UNIFORM ESTIMATES

It follows from the classical textbook [24] pp. 178-179 that for every nonnegative function $\bar u \in W(0,T,\Omega_x)$ (and $\bar h$ $s.t.$ $\bar h(t,x) = T_l(\bar u(t,x))$) there exists a solution
$u \in W(0,T,\Omega_x)$ of the parabolic problem with smooth coefficients

$$\frac{S_0}{2}\partial_t u - \frac{\tilde K}{2}\nabla'\cdot(\nabla' u) = -\chi_0(u)\int_{\bar h(t,x)}^{h_{soil}(x)} \big(\frac{\partial \bar p}{\partial t} - Q\big)\,dz, \tag{4.7}$$

$$\nabla u\cdot\vec\nu = 0 \quad \text{on} \quad (0,T)\times\partial\Omega_x, \quad u(0,x) = (h_0 - h_{bot})^2 \quad \text{in} \quad \Omega_x.$$

Multiplying Eq. (4.7) by u and integrating by parts over Ω_x, we thus obtain

$$\begin{aligned}\frac{S_0}{2}\frac{d}{dt}\int_{\Omega_x}|u(t,.)|^2\,dx + \frac{\tilde K}{2}\int_{\Omega_x}|\nabla u|^2\,dx \quad &\leq \quad |\langle \int_{\bar h(t,x)}^{h_{soil}} \frac{\partial \bar p}{\partial t} dz,\ u\rangle_{H^1(\Omega_x)',H^1(\Omega_x)}| \\ &\quad + |\int_{\Omega_x}\big(\int_{\bar h(t,x)}^{h_{soil}} Q\,dz\big)u\,dx|.\end{aligned} \tag{4.8}$$

Furthermore, by definition of the function T_l, we have

$$\begin{aligned}|\langle \int_{\bar h(t,x)}^{h_{soil}} \frac{\partial \bar p}{\partial t} dz,\ u\rangle_{H^1(\Omega_x)',H^1(\Omega_x)}| \quad &\leq \quad \|h_{soil} - h_{bot}\|_\infty^{1/2}\,\|u\|_{H^1(\Omega_x)}\|\partial_t \bar p\|_{(H^1(\Omega))'} \\ &\leq \quad \frac{\tilde K}{4}\big(\|u\|^2_{L^2(\Omega_x)} + \|\nabla u\|^2_{L^2(\Omega_x)}\big) + \tilde K\|h_{soil} - h_{bot}\|_\infty\|\partial_t \bar p\|^2_{(H^1(\Omega))'},\end{aligned}$$

$$\begin{aligned}|\int_{\Omega_x}\big(\int_{\bar h(t,x)}^{h_{soil}(x)} Q\,dz\big)u\,dx| \quad &\leq \quad \|u\|_{L^2(\Omega_x)}\|h_{soil} - h_{bot}\|_\infty^{1/2}\|Q\|_{L^2(\Omega)}, \\ &\leq \quad \frac{1}{2}\|u\|^2_{L^2(\Omega_x)} + \frac{1}{2}\|h_{soil} - h_{bot}\|_\infty\,\|Q\|^2_{L^2(\Omega)}.\end{aligned}$$

Applying Gronwall's inequality in its differential form, we get

$$\|u(t,.)\|_{L^2(\Omega_x)} \quad \leq \quad e^{\frac{(1+\tilde K/2)\,T}{S_0}}\big(\|u_0\|_{L^2(\Omega_x)} + \|h_{soil} - h_{bot}\|_\infty\big(\|Q\|^2_{L^2((0,T)\times\Omega)} + 2\,\tilde K\|\partial_t \bar p\|^2_{L^2(0,T,(H^1(\Omega))')}\big).$$

Then from this estimate and (4.8), we deduce

$$\| u\|^2_{L^2(0,T;\, H^1(\Omega_x))} \quad \leq \quad C(T, S_0, \tilde{K}, \|h_{soil} - h_{bot}\|_\infty, Q, C'_p) := C_u.$$

On the other hand

$$\begin{aligned}
&\|\frac{du}{dt}\|_{L^2(0,T;\, H^1(\Omega_x)')} = sup_{\|v\|_{L^2(0,T;\, H^1(\Omega_x))}\leq 1} |\int_0^T \langle \frac{du}{dt}, v\rangle_{H^1(\Omega_x)', H^1(\Omega_x))} dt| \\
&\leq \frac{2}{S_0}\Big(\frac{\tilde{K}}{2}\|u(t,.)\|_{L^2(0,T;H^1(\Omega_x))} + \|h_{soil} - h_{bot}\|_\infty\big(\|Q\|_{L^2((0,T)\times\Omega)} + \|\partial_t \bar{p}\|_{L^2(0,T,(H^1(\Omega))')}\big)\Big) \\
&\leq \frac{2}{S_0}\Big(\frac{\tilde{K}}{2}\, C_u \, + \|h_{soil} - h_{bot}\|_\infty\big(\|Q\|_{L^2((0,T)\times\Omega)} + \, C'_p\big)\Big) \, := C'_u.
\end{aligned}$$

STEP 2. NONNEGATIVITY OF THE SOLUTIONS.
Let us solely prove that $0 \leq u(t,x)$ for all $t \in (0,T)$ and for almost every $x \in \Omega_x$. Let $u_m = \sup(0, -u)$. The function u_m belongs to $L^2(0,T;V(\Omega_x))$, and is such that $\nabla u_m = -\chi_{\{u<0\}}\nabla u$ (see [12] Lemma 2.1; χ_A denotes the characteristic function of a set A). Let $\tau \in (0,T)$. Setting $w(t,x) = -u_m(x,t)\chi_{(0,\tau)}(t)$ in (4.7) results in

$$\begin{aligned}
&\frac{S_0}{2}\int_0^\tau \langle \partial_t u, -u_m\rangle_{V',V} + \frac{\tilde{K}}{2}\int_0^\tau\int_\Omega \chi_{\{u<0\}}|\nabla u|^2 \\
&= \langle \int_{\bar{h}(t,x)}^{h_{soil}} \frac{\partial \bar{p}}{\partial t} dz,\ \chi_0(u)\, u_m\rangle_{H^1(\Omega_x)', H^1(\Omega_x)} - \int_{\Omega_x} \big(\int_{\bar{h}(t,x)}^{h_{soil}} Q\, dz\big)\chi_0(u)\, u_m\, dx. \qquad (4.9)
\end{aligned}$$

In order to evaluate the first term in the left hand side of (4.9), we apply Lemma 2 with function f defined by $f(\lambda) = \max(0, -\lambda)$, $\lambda \in \mathbb{R}$. Of course $u_m(t,x) \neq 0$ iff $u(t,x) < 0$. We have

$$\int_0^\tau \langle \partial_t u, -u_m\rangle_{V',V} dt = \frac{1}{2}\int_\Omega \big(u_m^2(\tau,x) - u_m^2(0,x)\big)\, dx = \frac{1}{2}\int_\Omega u_m^2(\tau,x)\, dx.$$

Since $\chi_0(u)\chi_{\{u_1<0\}} = 0$ by definition of χ_0, the two last terms in the right hand side of (4.9) are null. Hence, (4.9) gives

$$\frac{S_0}{2}\int_\Omega u_m^2(\tau,x)\, dx \leq -\int_0^\tau\int_\Omega \frac{\tilde{K}}{2}\chi_{\{u<0\}}|\nabla u|^2\, dxdt \leq 0$$

and $u_m = 0$ a.e. in Ω_T.
STEP 3. UNIQUENESS
The uniqueness of the solution is obvious since the solution u is nonnegative. Indeed, if u_1 and u_2 are two solutions of (4.5)-(4.6), then $u = u_1 - u_2$ satisfies

$$\begin{aligned}
&\frac{S_0}{2}\partial_t u - \frac{\tilde{K}}{2}\nabla'\cdot(\nabla' u) = 0 \quad in \qquad (0,T)\times\Omega_x, \\
&\nabla u\cdot\vec{\nu} = 0 \quad \text{on} \quad (0,T)\times\partial\Omega_x, \quad u(0,x) = 0 \quad \text{in} \quad \Omega_x.
\end{aligned}$$

Following the previous computations, we infer from Gronwall lemma that $u = 0$ a.e. in $(0,T)\times\Omega_x$. This ends the proof of Lemma 4.3. □
The results stated in the lemma 3.1 require having regular non-cylindrical domains in particular with sufficiently regular boundaries (of class $\mathcal{C}^1$ by pieces as mentioned by Mignot). Since in our problem, we cannot guarantee as much regularity at the interface h (which is in $W(0,T,\Omega_x)$), we use a regularization process to place our study within the framework of Mignot [29].
We thus regularize h by convolution in space. Let $\psi \in C^\infty(\mathbb{R}^2)$, $\psi \geq 0$, with support in the unit ball such that $\int_{\mathbb{R}^2}\psi(x)dx = 1$. For $\eta > 0$ small enough, we set $\psi_\eta(x) = \psi(x/\eta)/\eta^2$. We extend h by zero outside Ω_x, so we have $h \in C([0,T];L^2(\mathbb{R}^2)) \cap W(0,T,\mathbb{R}^2)$. Hence we define $\tilde{h}$ by the convolution product with respect to the space variable

$$\tilde{h} = \psi_\eta * h.$$

Its restriction to Ω_x is denoted in the same way. It fulfills $\tilde{h} \in C^\infty(\bar{\Omega}_x)$, and as $\eta \to 0$, we have

$$\tilde{h} \to h \text{ strongly in } C([0,T];L^2(\Omega_x)) \cap L^2(0,T,H^1(\Omega_x)).$$

In Eqs. (4.1)-(4.4), we replace h by $\tilde{h}$ (the substitution appears in the space integration domain Ω_t).

Let $\bar{p} \in W_1$ and $\tilde{h}\,(=\psi_\eta * h) \in C^\infty(\bar{\Omega}_x)$ where h is given by Lemma 4.3.
We thus consider the following linearized and regularized problem in Ω_T : Find $p_\eta \in W(0,T,\Omega)$ s.t. $\forall \phi \in \mathcal{A}(\mathcal{O})$ (null on the interface Γ_t defined by $\tilde{h}$)

$$\langle \partial_t p_\eta, \phi\rangle_{F',F} + \int_0^T \Big(\int_{\Omega_t} \big(\tau(\bar{p})\nabla p_\eta + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar{p}))K_0\vec{e_3}\big)\cdot\nabla\phi\big) - Q\phi\Big)\, dt = 0, \tag{4.10}$$

$$p_\eta = 0 \text{ in } \Omega\setminus\Omega_t,\ \forall t\in[0,T] \text{ and } p_\eta(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0. \tag{4.11}$$

Proposition 4.4. *For any $\eta > 0$, there exists a unique function p_η in $W(0,T,\Omega)$ solution of (4.10)-(4.11). It fulfills the uniform estimates*

$$\|p_\eta\|_{L^2(0,T;H^1(\Omega))} \le C_p \quad \textit{and} \quad \|p_\eta\|_{L^2(0,T;(H^1(\Omega))')} \le C_p', \tag{4.12}$$

where C_p and C_p' only depend on the data of the original problem (3.8)- (3.9).

Let us admit for the moment this Proposition whose the proof will be given at the end. From now, we omit the subscript η in p_η (and then in u_η).

Let $(\bar{u},\bar{p}) \in W(0,T,\Omega)\times W(0,T,\Omega_x)$ the unique solution of (4.5)-(4.6) and (4.10)-(4.11), Lemma 4.3 and Proposition 4.4 enable to define an application $\mathcal{F}$ such that:

$$W(0,T,\Omega_x)\times W(0,T,\Omega) \to W(0,T,\Omega_x)\times W(0,T,\Omega)$$

$$\mathcal{F}(\bar{u},\bar{p}) = (u,p). \tag{4.13}$$

The end of the present subsection is devoted to the proof of the existence of a fixed point of $\mathcal{F}$ in some appropriate subset. We conclude the proof of Theorem 3.7 by passing to the limit when $\eta \to 0$.

Lemma 4.5. *Let $(\bar{u},\bar{p}) \in W(0,T,\Omega)\times W(0,T,\Omega_x)$ the unique solution of (4.5)-(4.6) and (4.10)-(4.11), thus*

- *There exists $\mathcal{C}$ a nonempty, closed, convex, bounded set in $W(0,T,\Omega_x)\times W(0,T,\Omega)$ satisfying $\mathcal{F}(\mathcal{C})\subset\mathcal{C}$,*
- *The application $\mathcal{F}$ defined by (4.13) is weakly sequentially continuous in $W(0,T,\Omega_x)\times W(0,T,\Omega)$.*

Proof.
We set $\mathcal{C} = W_1\times W_2$, the first point of Lemma 4.5 is obvious thanks to Lemma 4.3 and Proposition 4.4. Indeed $\mathcal{C}$ is clearly a nonempty (strongly) closed convex set in $W(0,T,\Omega_x)\times W(0,T,\Omega)$.
Regarding the second point of Lemma 4.5, we first note that $\mathcal{C}$ is compact for the weak topology. $\mathcal{F}$ maps $W_1\times W_2$ into it self. Let now $(v_n)_{n\ge0} = (\bar{u}_n,\bar{p}_n)_{n\ge0}$ be any sequence in $\mathcal{C}$ which is weakly convergent in $W(0,T,\Omega_x)\times W(0,T,\Omega)$, and let $v=(\bar{u},\bar{p})$ be its weak limit. We aim to show that

$$\mathcal{F}(v_n) \rightharpoonup \mathcal{F}(v) \quad \text{in } W(0,T,\Omega_x)\times W(0,T,\Omega) \text{ as } n\to\infty.$$

Since $\mathcal{F}(v_n)\in W_1\times W_2$ and $W_1\times W_2$ is weakly compact, it is sufficient to show that there exits a subsequence (v_n') of (v_n) such that $\mathcal{F}(v_n') \rightharpoonup \mathcal{F}(v)$. Extracting a subsequence if needed we may assume without loss of generality that $\mathcal{F}(v_n)\rightharpoonup w$ in $W(0,T,\Omega_x)\times W(0,T,\Omega)$ as $n\to\infty$ for some $w=(u,p)\in W_1\times W_2$, and we have to show that w and $\mathcal{F}(v)$ agree. Set $w_n = \mathcal{F}(v_n)$ $(w_n=(u_n,p_n))$, it follows from Aubin's Lemma that

$$\begin{aligned}
w_n \to w \text{ in } L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega) \text{ and} \quad & w_n(t,x)\to w(t,x) \text{ a.e.;}\\
v_n \to v \text{ in } L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega) \text{ and} \quad & v_n(t,x)\to v(t,x) \text{ a.e.;}\\
\partial_t w_n \rightharpoonup \partial_t w \qquad \text{in} \quad & L^2(0,T;H^1(\Omega_x)')\times L^2(0,T;H^1(\Omega)')\\
\nabla w_n \rightharpoonup \nabla w \qquad \text{weakly in} \quad & L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega).
\end{aligned}$$

Thanks to Lebesgue theorem (and the properties of functions τ and T_l) we obtain that $w = \mathcal{F}(v)$ (since $w(0,.) = (u(0,.),p(0,.)) = (u_0,p_0)$ because $w\in\mathcal{C}$) and the proof that $\mathcal{F}|_{\mathcal{C}}$ be weakly sequentially continuous is complete.
It follows from Schauder theorem [39] that there exists $(u,p)\in W_1\times W_2$ such that $\mathcal{F}(u,p)=(u,p)$. The proof of Lemma 4.5 is thus achieved. □

We collect the results obtained previously. We can associate with any real number $\eta > 0$ the fixed point point $(u_\eta, p_\eta) \in W_1 \times W_2$ of the mapping $\mathcal{F}$. It is a solution of the system :

$$\partial_t p_\eta - \nabla \cdot \big(\tau(p_\eta)\nabla p_\eta\big) - \frac{\rho_0\, g}{\mu}\nabla \cdot \Big(\kappa(\mathcal{P}^{-1}(p_\eta))K_0\vec{e_3}\Big) = Q \quad \text{in} \quad \mathcal{O}_T, \tag{4.14}$$

$$p_\eta\big|_{\Gamma_t} = \mathcal{P}(P_s) \quad \text{in} \quad (0,T), \qquad \nabla\big(\mathcal{P}^{-1}(p_\eta) + \rho_0 g z\big)\cdot\vec{\nu} = 0 \quad \text{on} \quad (0,T)\times(\Gamma_{soil}\cup\Gamma_{ver}),$$

$$p_\eta(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0. \tag{4.15}$$

$$\frac{S_0}{2}\partial_t u_\eta - \frac{\tilde{K}}{2}\nabla'\cdot(\nabla' u_\eta) = -\chi_0(u_\eta)\int_{h_\eta(t,x)}^{h_{soil}(x)}\big(\frac{\partial p_\eta}{\partial t} - Q\big)\,dz \quad in \qquad (0,T)\times\Omega_x, \tag{4.16}$$

$$\nabla u_\eta\cdot\vec{\nu} = 0 \quad \text{on} \quad (0,T)\times\partial\Omega_x, \quad u_\eta(0,x) = (h_0(x) - h_{bot}(x))^2 \quad \text{in} \quad \Omega_x. \tag{4.17}$$

We can obtain similar estimates for (u_η, p_η) than those derived in Lemma 4.3 and Proposition 4.4. We thus assert the existence of limit functions (extracting a subsequence if needed) $(u,p) \in W(0,T,\Omega_x)\times W(0,T,\Omega)$ such that

$$\begin{aligned}
(u_\eta, p_\eta) &\to (u,p) && \text{in} && L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega)\\
(u_\eta(t,x), p_\eta(t,x)) &\to (u(t,x), p(t,x)) && \text{a.e in} && ((0,T)\times\Omega_x)\times((0,T)\times\Omega)\\
\tilde{h}(t,x) = \psi_\eta * h(t,x) &\to h(t,x), && \text{a.e in} && (0,T)\times\Omega_x\\
(\partial_t u_\eta, \partial_t p_\eta) &\rightharpoonup (\partial_t u, \partial_t p) && \text{in} && L^2(0,T;H^1(\Omega_x)')\times L^2(0,T;H^1(\Omega)')\\
(\nabla u_\eta, \nabla p_\eta) &\rightharpoonup (\nabla u, \nabla p) && \text{weakly in} && L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega).
\end{aligned}$$

Letting $\eta \to 0$ in weak formulations resulting from (4.14)-(4.17), we prove the existence of a weak solution (u,p) of problem (3.8)-(3.11). This ends the proof of Theorem 3.7. □

4.2. Proof of Proposition 4.4

Again, we omit the subscript η in p_η.
The proof of Proposition 4.4 is done in two steps. We first use the method of auxiliary domains presented in [29] to solve difficulties related to the free boundary. That is, we extend the functions out of the domain of study by zero, then we introduce a penalized problem and go to the limit to return to the linearized problem (4.10)-(4.11).
We thus consider the weak solution p of the linearized problem (4.10)-(4.11).
So, $\forall\phi \in L^2([0,T], H^1(\Omega_t))(= H^{0,1}(\mathcal{O}_T))$ with $\phi\big|_{\Gamma_t} = 0$, we look for $p \in W(0,T,\Omega)$ such that

$$\langle\partial_t p, \phi\rangle_{F',F} + \int_0^T\Big(\int_{\Omega_t}\big(\tau(\bar{p})\nabla p + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar{p}))K_0\vec{e_3}\big)\cdot\nabla\phi\big) - Q\phi\Big)\,dt = 0.$$

We first remark that the solution of system (4.10)-(4.11) is unique. Indeed, if p_1 and p_2 are two solutions of (4.10)-(4.11), then $q = p_1 - p_2$ satisfies

$$\langle\partial_t q, \phi\rangle_{F',F} + \int_0^T\Big(\int_{\Omega_t}\big(\tau(\bar{p})\nabla q\cdot\nabla\phi\big)\Big)\,dt = 0.$$

Then, taking $\phi = q$ and using the fourth point of Lemma 3.1, we conclude that

$$\frac{1}{2}\int_{\Omega_T} q^2(T,x)\;d\mathbf{x}\,dt + \int_0^T\int_{\Omega_t}\tau(\bar{p})|\nabla q|^2\;d\mathbf{x}\,dt = 0,$$

since $q(0,.) = 0$. We infer from this equality that $q = 0$ a.e. in $(0,T)\times\Omega$ (since $q = 0$ on the interface Γ_T).
We will define a family of approximate problems which are linear parabolic problems in the cylindrical domain $(0,T)\times\Omega$, and whose the solution restricted to the Ω_T set will converge to the p solution of the linearized equation (4.10).

Step 1. Penalized problems

Let $\epsilon > 0$, we now consider the following penalized problem on Ω : Find $p_\epsilon \in W(0,T,\Omega)$ s.t. $\forall \phi \in L^2(0,T;\mathcal{D}(\Omega))$

$$\langle \partial_t p_\epsilon, \phi\rangle_{F',F} + \int_0^T \Big(\int_\Omega \big(\tilde\tau(\bar p)\nabla p_\epsilon + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))\tilde K_0 \vec{e_3}\big)\cdot\nabla\phi\Big) - \tilde Q\phi\Big)\, d\mathbf{x}\, dt$$
$$+ \int_{\mathcal{O}'_T} \nabla p_\epsilon\cdot\nabla\phi\, d\mathbf{x}\, dt + \frac{1}{\epsilon}\int_{\mathcal{O}'_T} p_\epsilon\phi\, d\mathbf{x}\, dt = 0, \tag{4.18}$$
$$p_\epsilon(0,x,z) = \mathcal{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0 \quad \text{and} \quad p_\epsilon(0,x,z) = 0 \quad \text{in} \quad \Omega\setminus\Omega_0. \tag{4.19}$$

where $\tilde K_0 = K_0$ in $\mathcal{O}_T$ and $\tilde K_0 = 0$ in $\mathcal{O}'_T$, $\tau(p) = \dfrac{\tilde K_0}{\mu}\dfrac{\kappa(\mathcal{P}^{-1}(p))}{(\theta' + \alpha_P\theta)(\mathcal{P}^{-1}(p))}$ and
$\tilde Q = Q$ in $\mathcal{O}_T$ and $\tilde Q = 0$ in $\mathcal{O}'_T$.
We aim to state that the penalized system (4.18)-(4.19) admits a unique solution p_ϵ which tends to the solution of problem (4.10)-(4.11) when $\epsilon \to 0$. Eq. (4.18) can be written

$$\langle \partial_t p_\epsilon, \phi\rangle_{F',F} + \underbrace{\int_0^T\int_\Omega \tilde\tau(\bar p)\nabla p_\epsilon\cdot\nabla\phi\, d\mathbf{x}\, dt + \int_{\mathcal{O}'_T}\nabla p_\epsilon\cdot\nabla\phi\, d\mathbf{x}\, dt + \frac{1}{\epsilon}\int_{\mathcal{O}'_T} p_\epsilon\phi\, d\mathbf{x}\, dt}_{A_\epsilon(,p_\epsilon,\phi)}$$
$$= \underbrace{-\int_0^T\int_\Omega \Big(\frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))\tilde K_0\vec{e_3}\cdot\nabla\phi\Big) - \tilde Q\phi\Big)\, d\mathbf{x}\, dt}_{L_\epsilon(\phi)} \qquad \forall\phi\in W(0,T,\Omega). \tag{4.20}$$

Due to (3.5), we establish that the coefficients of A_ϵ are in $L^\infty((0,T)\times\Omega)$. Moreover, we have

$$A_\epsilon(p,p) \geq inf(1,\tau_-,\frac{1}{\epsilon})\,\|p\|_{L^2(0,T,H^1(\Omega))}, \quad \forall p\in L^2(0,T,H^1(\Omega)).$$

We directly check that L_ϵ is a linear form on $L^2(0,T,H^1(\Omega))$. We thus deduce the existence and uniqueness for the system (4.18)-(4.19).

STEP 2. LIMIT WHEN $\epsilon \to 0$

We first derive some uniforme estimates with respect to ϵ (and also η). Multiplying Eq. (4.21) by p_ϵ and integrating by parts over Ω, we thus obtain $\forall s \leq T$

$$\langle \partial_t p_\epsilon, p_\epsilon\rangle_{F',F} + \underbrace{\int_0^s\int_{\Omega_t}\tilde\tau(\bar p)|\nabla p_\epsilon|^2\, d\mathbf{x}\, dt + \int_{\mathcal{O}'_s}|\nabla p_\epsilon|^2\, d\mathbf{x}\, dt + \frac{1}{\epsilon}\int_{\mathcal{O}'_s} p_\epsilon^2\, d\mathbf{x}\, dt}_{I_1}$$
$$= \underbrace{-\int_0^s\int_\Omega \Big(\tilde Q p_\epsilon + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))\tilde K_0\vec{e_3}\cdot\nabla p_\epsilon\Big)\, d\mathbf{x}\, dt}_{I_2} \qquad \forall\phi\in W(0,T,\Omega). \tag{4.21}$$

Then, applying Lemma 3.1 to the first term, we get

$$|\langle \partial_t p_\epsilon, p_\epsilon\rangle_{F',F}| = |\int_0^s \langle \partial_t p_\epsilon, p_\epsilon\rangle_{V'(\Omega),V(\Omega)}\, dt| \quad = \quad \frac{1}{2}\Big(\int_\Omega p_\epsilon^2(s,.)d\mathbf{x} - \int_\Omega p_\epsilon^2(0,.)d\mathbf{x}\Big)$$

Besides

$$|I_1| \quad \geq \quad \tau_-\,\|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega_t))} + \|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega'_t))} + \frac{1}{\epsilon}\int_{\mathcal{O}'_s} p_\epsilon^2\, d\mathbf{x}\, dt,$$
$$|I_2| \quad \leq \quad \frac{\epsilon_1}{2}\int_0^s\int_\Omega p_\epsilon^2(s,.)d\mathbf{x} + \frac{2}{\epsilon_1}\|Q\|^2_{L^2([0,T],L^2(\Omega_t))} + \frac{\epsilon_2}{2}\,\|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega_t))}$$
$$+ \frac{2}{\epsilon_2}\,T\,mes(\Omega)\Big(\frac{\rho_0\, g\kappa_+ K_0}{\mu}\Big)^2.$$

By taking $\epsilon_1 = 1/2$ and $\epsilon_2 = \tau_-$, we deduce directly from these estimates that the sequence $\{p_\epsilon\}$ is bounded in $L^2(0,T;H^1(\Omega))$ and the sequence $\{\frac{1}{\sqrt{\epsilon}}p_\epsilon\}$ is bounded in $L^2((0,T)\times\Omega'_T)$. Thanks to Gronwall's Lemma, we deduce that there exists a constant C_p depending only on the data such that

$$\|p_\epsilon\|^2_{L^2([0,T],H1(\Omega_t))} \quad \leq \quad C_p \tag{4.22}$$

In particular we have

$$\|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega_t))} \quad \leq \quad \frac{C_0}{\tau_-}\left(1+\frac{\tau_- T}{2}e^{\frac{\tau_- T}{2}}\right),$$

where

$$C_0 \;=\; \int_\Omega p_0^2 d\mathbf{x} \;+ 8\,\|Q\|^2_{L^2([0,T],L^2(\Omega_t))} + \frac{4}{\tau_-}\,T\,mes(\Omega)\Big(\frac{\rho_0\, g\kappa_+ K_0}{\mu}\Big)^2.$$

We can thus extract subsequences $\{p_\epsilon\}$, $\{\frac{1}{\sqrt{\epsilon}}p_\epsilon\}$ (not relabeled for convenience) and there exist $q \in L^2(0,T;\, H^1(\Omega))$ and $q' \in L^2((0,T)\times\Omega'_T)$ such that

$$p_\epsilon \rightharpoonup q \quad \text{weakly in} \quad L^2((0,T)\times\Omega) \tag{4.23}$$

$$\frac{1}{\sqrt{\epsilon}}p_\epsilon \rightharpoonup q' \quad \text{weakly in} \quad L^2((0,T)\times\Omega'_T) \tag{4.24}$$

$$\nabla p_\epsilon \rightharpoonup \nabla p \quad \text{weakly in} \quad L^2((0,T)\times\Omega). \tag{4.25}$$

It results from the first two convergences that

$$p_\epsilon|_{\mathcal{O}'_T} \rightharpoonup q|_{\mathcal{O}'_T} \quad \text{weakly in} \quad L^2((0,T)\times\Omega'_T)$$

$$p_\epsilon|_{\mathcal{O}'_T} = \sqrt{\epsilon}\times\frac{1}{\sqrt{\epsilon}}p_\epsilon \rightharpoonup 0 \quad \text{weakly in} \quad L^2((0,T)\times\Omega'_T),$$

so

$$q|_{\mathcal{O}'_T} = 0. \tag{4.26}$$

Moreover, since $q \in F([0,T]\times\Omega)$, we infer from the second result of Lemma 3.1, that we can define $\gamma(q)$ on Γ_t and $\gamma(q)=0$ on Γ_t, $0\leq t\leq T$ thanks to (4.26), and thus $q|_{\mathcal{O}_T} \in F(\mathcal{O}_T)$.

We must now check that $q|_{\mathcal{O}_T}$ satisfies (4.10). Let $\phi \in F(\mathcal{O}_T)$ that we extend by zero on $\mathcal{O}'_T$, the extension is still denoted by ϕ. Thus taking $\phi \in F((0,T)\times\Omega)$ in (4.21), we get

$$\int_0^T \langle \partial_t p_\epsilon, \phi\rangle_{V',V}\, dt + \int_0^T \Big(\int_\Omega \big(\tau(\bar p)\nabla p_\epsilon + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))K_0\vec{e_3}\big)\cdot\nabla\phi\big) - Q\phi\Big)\, d\mathbf{x}\, dt = 0,$$

that we can write as follows by choosing $\phi \in B_T(\mathcal{O}_T)$

$$\begin{aligned} -\int_0^T \langle p_\epsilon, \partial_t\phi\rangle_{V',V}\, dt \;\;+\;\; & \int_0^T \Big(\int_\Omega \big(\tau(\bar p)\nabla p_\epsilon + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))K_0\vec{e_3}\big)\cdot\nabla\phi\big) - Q\phi\Big)\, d\mathbf{x}\, dt \\ = \;& \int_{\Omega_0} p_0(\mathbf{x})\phi_0(0,\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

By letting $\epsilon\to 0$, we obtain

$$\begin{aligned} -\int_0^T \langle q|\mathcal{O}_T, \partial_t\phi\rangle_{V',V}\, dt \;\;+\;\; & \int_0^T \Big(\int_{\Omega_t} \big(\tau(\bar p)\nabla q|\mathcal{O}_T + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))K_0\vec{e_3}\big)\cdot\nabla\phi\big) - Q\phi\Big)\, d\mathbf{x}\, dt \\ = \;& \int_{\Omega_0} p_0(\mathbf{x})\phi_0(0,\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{4.27}$$

It remains to be established that

$$D_t(q|\mathcal{O}_T) \in F'(\mathcal{O}_T) \quad \text{and} \quad D_t(p_\epsilon|\mathcal{O}_T) \rightharpoonup D_t(q|\mathcal{O}_T) \quad \text{in} \quad F'(\mathcal{O}_T).$$

From (4.27), we deduce that, taking $\phi \in \mathcal{D}(\mathcal{O}_T)$

$$D_t(q|\mathcal{O}_T) - \nabla\cdot\big(\tau(\bar p)\nabla q|\mathcal{O}_T + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))K_0\vec{e_3}\big) = Q \quad \text{in} \quad \mathcal{D}'(\mathcal{O}_T).$$

Since $D'(\mathcal{O}_T)$ is dense in $F'(\mathcal{O}_T)$, the previous equality holds true in $F'(\mathcal{O}_T)$. Moreover the solution $p_\epsilon|\mathcal{O}_T$ of (4.21) verifies

$$D_t(p_\epsilon|\mathcal{O}_T) - \nabla\cdot\big(\tau(\bar p)\nabla p_\epsilon|\mathcal{O}_T + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))K_0\vec{e_3}\big) = Q \quad \text{in} \quad F'(\mathcal{O}_T).$$

Letting $\epsilon\to 0$ and we infer from convergences (4.23)-(4.25) that

$$\underbrace{\nabla\cdot\big(\tau(\bar p)\nabla p_\epsilon|\mathcal{O}_T + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))K_0\vec{e_3}\big) - Q}_{=D_t(p_\epsilon|\mathcal{O}_T)} \rightharpoonup \underbrace{\nabla\cdot\big(\tau(\bar p)\nabla q|\mathcal{O}_T + \frac{\rho_0\, g}{\mu}\kappa(\mathcal{P}^{-1}(\bar p))K_0\vec{e_3}\big) - Q}_{=D_t(q|\mathcal{O}_T)},$$

weakly in $F'(\mathcal{O}_T)$. Thus $q|\mathcal{O}_T$ is the unique solution of (4.10)-(4.11), and the limit of $p_\epsilon|_{\mathcal{O}_T}$ being independent of the chosen subsequence, the whole sequence converges towards $q|\mathcal{O}_T$ in $\mathcal{A}(\mathcal{O}_T)$. Moreover, we obtain the first part of (4.12) for the solution $q \in L^2(0,T,H^1(\Omega))$ of system (4.10)-(4.11) in the same way that for estimate (4.22) obtained for p_ϵ. Finally, as was done in Lemma 4.3, we deduce from the first inequality of (4.12) that

$$\|\partial_t q\|^2_{L^2(0,T,H^1(\Omega)')} \quad \leq \quad C'_p,$$

where C'_p depends on the data and on C_p. This ends the proof of Proposition 4.4. □

References

[1] MB Abbott, JC Bathurst, JA Cunge, PE O'connell, and J Rasmussen. An introduction to the european hydrological system - systeme hydrologique europeen,"she", 2: Structure of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1):61–77, 1986.

[2] P. Ackerer, A. Younes, *Efficient approximations for the the simulation of density driven flow in porous media*, Adv. Water Res., Vol. 31, 15–27, 2008.

[3] Alkhayal, J., Issa, S., Jazar, M., Monneau, R.: Existence results for degenerate cross-diffusion systems with application to seawater intrusion, ESAIM Control Optim. Calc. Var. 24, no. 4,1735-1758 (2018)

[4] H. W. Alt, S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., Vol. 1, 311–341, 1983.

[5] HW Alt and E. Di Benedetto, Nonsteady flow of water and oil through inhomogeneous porous media, Ann. Scuola Norm. Sup. Pisa 12 (1985) 335-392.

[6] A. Bensoussan, J. L. Lion, G. Papanicoulou, Asymptotic analysis for periodic structure, North-Holland, Amsterdam, 1978.

[7] Christine Bernardi, Adel Blouza, and Linda El Alaoui. The rain on underground porous media part i: Analysis of a richards model. *Chinese Annals of Mathematics, Series B*, 34(2):193–212, Mar 2013.

[8] Heiko Berninger, Mario Ohlberger, Oliver Sander, and Kathrin Smetana. Unsaturated subsurface flow with surface water and nonlinear in- and outflow conditions. *Mathematical Models and Methods in Applied Sciences*, 24(05):901–936, 2014.

[9] R.H. Brooks and A.T. Corey. *Hydraulic Properties of Porous Media*. Colorado State University Hydrology Papers. Colorado State University, 1964.

[10] Bear Jacob. *Dynamics of fluids in porous media*. Elsevier, New-York, 1972.

[11] Bear Jacob and Verruijt Arnold. *Modeling groundwater flow and pollution*. Springer, Netherlands, 1987.

[12] P. Benilan, L. Boccardo, T. Gallouët, R. Gariepy, M. Pierre, J.L. Vazquez, *An L^1 theory of existence and uniqueness of nonlinear elliptic equations*, Ann. Sc. Norm. Super. Pisa, Cl. Sci., IV. Ser., 22 (1995), 240–273.

[13] C. Bourel, C. Choquet, C. Rosier and M. Tsegmid, *Modelling of shallow aquifers in interaction with overland water*, submitted.

[14] X. Chen, A. Friedman and T Kimura, Nonstationary filtration in partially saturated porous medium, Euro. J. Appl. Math. 5 (1994) 405-429.

[15] C. Choquet,, M. M. Diédhiou, C. Rosier, *Derivation of a Sharp-Diffuse Interfaces Model for Seawater Intrusion in a Free Aquifer. Numerical Simulations*, SIAM J. Appl. Math. 76 (2016), no. 1, 138-158.

[16] H. Darcy, *Les fontaines publiques de la ville de Dijon; exposition et application des principes à employer dans les questions de distribution d'eau*, Victor Dalmont, Editeur, Paris, (1856).

[17] , J. Dupuit1863, *Etudes thdoriques et pratiques sur le mouvement des eaux dans les canaux dcouverts et à travers les terrains permśables*, 2ème édition, Dunod, Paris, (1863).

[18] P. Sochala, A. Ern, and S. Piperno. Mass conservative bdf-discontinuous galerkin/explicit finite volume schemes for coupling subsurface and overland flows. *Computer Methods in Applied Mechanics and Engineering*, 198(27):2122 – 2136, 2009.

[19] L. C. Evans, Partial differential equations, American Mathematical Society, 1998.

[20] C.W. Fetter, *Hydrogeology: A shot history, Part* 2, Ground Water, 42 (2004), 949-953.

[21] J. Hulshof and N. Wolanski, Monotone flows in N-dimensional partially saturated porous media: Lipschitz continuity of the interface, Arch. Rat. Mech. Anal. 102 (1988) 287-305.

[22] G. Gagneux, M. Madaune-Tort, *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière*. Mathématiques & Applications, 22, Springer, 1996.

[23] Jun Kong, Pei Xin, Zhi yao Song, and Ling Li. A new model for coupling surface and subsurface water flows: With an application to a lagoon. *Journal of Hydrology*, 390(1):116 – 120, 2010.

[24] O.A. Ladyzhenskaja, V. A. Solonnikov and N. N. Ural'ceva, *Linear and quasilinear equations of parabolic type* (Amer. Math. Soc. 1968).

[25] J. L. Lions, *Sur les problèmes mixtes pour certains systèmes paraboliques dans des ouverts non cylindriques*; Annales de l'Institut Fourier 7, 1957, p. 143-182.

[26] J. L. Lions, E. Magenes, *Problèmes aux limites non homogènes*, Vol. 1, Dunod, 1968.

[27] C.M. Marle, *Henry Darcy et les 'ecoulements de fluides en milieux poreux*, Oil and Gas science and technology Re. IFP, Vol. 61 (5) (2006), 599-609.

[28] N. G. Meyers, An L^p-estimate for the gradient of solution of second order elliptic divergence equations, Ann. Sc. Norm. Sup. Pisa, Vol. 17, pp. 189-206, 1963.

[29] A. L. Mignot, *Méthodes d'approximation des solutions de certains problèmes aux limites linéaires*, Rendiconti del Seminario Matematico della Università di Padova, tome 40 (1968), p. 1-138.

[30] Hung Q Pham, Delwyn G Fredlund, and S Lee Barbour. A study of hysteresis models for soil-water characteristic curves. *Canadian Geotechnical Journal*, 42(6):1548–1568, 2005.

[31] Mary F Pikul, Robert L Street, and Irwin Remson. A numerical model based on coupled one-dimensional richards and boussinesq equations. *Water Resources Research*, 10(2):295–302, 1974.

[32] C. Rosier, L. Rosier, *Well-posedness of a degenerate parabolic equation issuing from two-dimensional perfect fluid dynamics*, Applicable Anal., Vol. 75 (3-4), pp 441–465, 2000.
[33] RE Showalter and N Su, Partially saturated flow in a poroelastic medium, Disc. Cont. Dyn. Syst. Ser. B (2001) 403-420.
[34] Ben Schweizer. Hysteresis in porous media: Modelling and analysis. *Interfaces and Free Boundaries*, 19:417–447, 01 2017.
[35] J. Simon, Compact sets in the space $L^p(0,T,B)$, Ann. Mat. Pura Appl., vol. 146 (4), 65–96, 1987.
[36] M. Tsegmid,*Modélisation d?aquifères peu profonds en interaction avec les eaux de surfaces*, Phd Thesis, 2019 ULCO. www.theses.fr/2019DUNK0526
[37] M Th Van Genuchten. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils 1. *Soil science society of America journal*, 44(5):892–898, 1980.
[38] H. M. Yin, *A singular-degenerate free boundary problem arising from the moisture evaporation in a partially saturated porous medium*, Ann. Mat. Pura Appl. **161** (1992) 379-397.
[39] E. Zeidler, *Nonlinear functional analysis and its applications*, Part 1, Springer Verlag, 1986.

[a] Univ. du Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France. [b] CNRS FR 2037, France. [c] Université Libanaise, LaMA-Liban, Laboratoire de recherche en Mathématiques et Applications, P.O. Box 37 Tripoli, Liban.
Email address: Safaa.Al-Nazer@etu.univ-littoral.fr

[a] Univ. du Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France.[b] CNRS FR 2037, France.
Email address: Carole.Rosier@univ-littoral.fr

[a] Univ. Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France. [b] CNRS FR 2037, France

Chapitre 4

Modèle couplant le flux de Richards 3D avec le flux horizontal de l'écoulement : Cas non isotrope et conservatif

GLOBAL EXISTENCE RESULT FOR A COUPLED SYSTEM DESCRIBING THE EXCHANGES BETWEEN A SHALLOW AQUIFER AND THE OVERLAND WATER.

SAFAA AL NAZER, CHRISTOPHE BOUREL, AND CAROLE ROSIER

Abstract. In this work, we analyze a new model which describes the water flow in shallow aquifers. It is an alternative to the 3D-Richards model which is classically used in this kind of porous media. Its derivation is guided by two ambitions: any new model should be low cost in computational time and should still give relevant results at every time scale. We thus keep track of two types of flow occurring in such a context and which are dominant when the *ratio* thickness over longitudinal length is small: the first one is dominant in a small time scale and is described by a vertical 1D-Richards problem; the second one corresponds to a large time scale, when the evolution of the hydraulic head turns to become independent of the vertical variable. The final model consists of a strongly coupled system of parabolic-type pde's that can degenerate according to the degeneration of the moisture content. We show how taking into account the low compressibility of the fluid makes it possible to improve the mathematical analysis and to give a result of existence of weak solutions in non-degenerate case.

Keywords: Richards equation; quasilinear parabolic equations; global in time existence.

1. Introduction

In the present paper, we focus on the hydrogeological aspect in aquifers, namely exchanges between the overland and the underground waters. We thus consider the displacement of a wetting phase (water) in the presence of a non-wetting fluid (air) in a porous medium. Assuming that the air present in the unsaturated zone has infinite mobility allows to use a model for immiscible fluid flow simplified by the Richards hypothesis. The saturation is thus considered as a monotone function depending of the pressure head and the so-called Richards model consists in a nonlinear three-dimensional equation of degenerate parabolic type. All the existing simplified models for the fluid displacement in aquifers are motivated by the characteristics of the flow in their saturated part. A form of stratification enables the definition of interfaces and the slowness of the natural dynamics ensures that these interfaces have a smooth and stable behaviour. Moreover the flows are essentially orthogonal to the walls (Dupuit's hypothesis). These points allow the vertical integration of the Richards equation in the saturated area and lead to the use of a family of 2D models developed since the 60's (see e.g. the works of Jacob Bear, [4, 5]). A main weakness of the approach by vertical integration lies in its justification. It is only valuable for very precise length and time scales, the time scale in particular being completely different of the typical durations of chemical reactions (see once again [4] for empirical and qualitative arguments, see [18] for asymptotic computations). However, such 2D models are widely used, even out of their validity range and even if it turns out to be especially difficult to properly couple them with the flow in the unsaturated part of the underground.

So we consider models given in [7] exploiting the low thickness of a confined or unconfined aquifer. They consist in capturing very different physical phenomena, the fast and essentially vertical leakage coming from the surface through an unsaturated soil and the slow and essentially horizontal displacement in the saturated part of the aquifer, that are classically modeled by mathematical systems with very different structures. Clearly, given its construction, the model is simpler to manipulate numerically since the original $3D$ problem is replaced by the coupling of a $2D$ problem with several independent $1D$-problems (which can be solved in parallel). This implies significant time savings in the numerical processing. Let $\epsilon > 0$ describe the ratio of the aquifer's deepness over its characteristic horizontal length. Assume that ϵ is small. The usual approach would consist in choosing a reference time for the study, introducing an asymptotic expansion of the solution of the 3D-Richards system and using the scale separation for identifying the equations governing the main order terms of this ansatz. This is the classical process for deriving an effective model. Here the asymptotic analysis is not used for deriving an effective model for a given reference time. Rather, it is used for proving that each model of our new class and the 3D-Richards equation are associated with the same effective problem for any time scale. Basically:

(1) At short times, the horizontal flow is very small and the vertical one satisfies a $1D$-Richards problem.

Date: October 31, 2020.

(2) At non-short times, the vertical flow appears instantaneous. The corresponding pressure profile satisfies the stationary $1D$-Richards problem. Then the hydraulic head H does not depend on the vertical variable z. This corresponds to the so-called Dupuit hypothesis.
(3) At large times, the horizontal flux is non-zero. It is ruled by a $2D$-horizontal diffusion equation where the conductivity is the vertical average of the permeability tensor on the *whole* depth of the aquifer.

The mathematical study of the model is particularly delicate because of nonlinearities, the free boundary between each area and the difficulty resulting from the coupling between the two zones which is expressed by terms of flux at the interface. We must also deal with the mathematical difficulties inherent in Richards equations. Finally, there is a general mathematical difficulty in the structure of the set of PDEs modeling the dynamics of underground water. Indeed, when considering a free water table, we must face the gradual disappearance of water in the desaturation zone and thus the disappearance of a main unknown of the problem. There exists a huge literature regarding the classical Richards equations. Let us mention the works of Alt *et al* ([3, 2]) and the papers [8, 11, 19] devoted to the study of the degenerate in time equation

$$\partial_t \theta(p) - \Delta p = 0,$$

where $\theta(p)$ denotes the moisture content. We quote also in the one-dimensional case the work of Yin ([21]) concerning the existence of weak solution for the fully degenerate problem

$$\partial_t \theta(p) - \partial_x(\kappa(\theta(p))\partial_x p) = 0,$$

when just assuming that $\theta', \kappa' > 0$.
Classically, the Kirchoff transform is applied to the Richards equation (under appropriate assumptions about porosity and permeability) to eliminate nonlinearity in the diffusive term. In this work, we instead exploit the hypothesis of low compressibility of water to eliminate the degeneracy in the time derivative term of the Richards equation.
This transformation brings us back to the framework of quasilinear parabolic equations on non cylindrical domains. We can thus apply the auxiliary domain method introduced by Lions and Mignot [14, 16] to deal with the free boundary.
Besides, the second equation governing the interface evolution consists in a nonlinear parabolic equation. Taking into account the compressibility of the water introduces a dependance with respect to the thickness of the saturated zone in the time derivative term.

The document is organized as follows:

- we continue this section by describing the geometry of the aquifer and by recalling the model presented in [7]. This model couple the fast and slow components of the flow that are dominant in shallow aquifers. The justification of the existence of the solution of this kind of problem is the main subject of this article. We end this section by explaining the obstacles we face in the theoretical analysis of this problem. In particular, one of them is the degeneracy in the time derivative term in the PDE changing its type from parbolic to elliptic. This difficulty is classical in the study of Richards equations. Our strategy to avoid this difficulty is the consideration of the small compressibility of the fluid: this is the first contribution of this article.
- In Section 2 we present a new coupled model more adapted to the mathematical analysis. This model is, from a physical point of view, very close to the previous one and differs in two ways:
  - It takes into account a weeak compressibility of the fluid. In particular, we use the same extention of the classical 3d-Richards equations to the case of a slightly compressible fluid as it is done in [17].
  - It considers a small horizontal conductivity instead of a vanishing one in the upper part of the aquifer.
- The main result is given in Section 3. It concerns the global in time existence of the solution of the model described in Section 2.
- The proof of the Theorem is performed in section 4. It consists in a fixed point strategy in order to deal with the difficulties linked to nonlinearities and coupling. The last subsection is devoted to the proof of the existence of an unique solution to the linearized problem with a free boundary by reducing it to a problem in a fixed domain.

1.1. Geometry, physical parameters and boundary conditions

For the three-dimensional description, we denote by $\boldsymbol{x} := (x, z)$, $x = (x_1, x_2) \in \mathbb{R}^2$, $z \in \mathbb{R}$, the usual coordinates.

Geometry. The aquifer is represented by a three-dimensional domain $\Omega := \Omega_x \times (h_{\text{bot}}, h_{\text{soil}})$, $\Omega_x \subset \mathbb{R}^n$ with $n \geq 2$, function h_{bot} (respect. h_{soil}) describing its lower (respect. upper) topography. The upper and lower surfaces are thus defined by the graph of the functions $h_{\text{bot}} = h_{\text{bot}}(x)$ and $h_{\text{soil}} = h_{\text{soil}}(x)$, $x \in \Omega_x$. We assume that

$$h_{\text{soil}}(x) > h_{\text{bot}}(x), \qquad \forall\, x \in \Omega_x. \tag{1.1}$$

More precisely the domain is given by:

$$\Omega = \Big\{(x,z) \in \Omega_x \times \mathbb{R} \quad | \quad z \in \big]\, h_{\text{bot}}(x), h_{\text{soil}}(x) \big[\Big\}. \tag{1.2}$$

We always denote by $\vec{\nu}$ the outward unit normal and $\vec{e_3}$ is the unitary vertical vector pointing up. We decompose the boundary $\partial\Omega$ of Ω in three zones (bottom, top and vertical)

$$\partial\Omega = \Gamma_{\text{bot}} \sqcup \Gamma_{\text{soil}} \sqcup \Gamma_{\text{ver}}\,,$$

with

$$\Gamma_{\text{bot}} := \Big\{(x,z) \in \Omega \mid z = h_{\text{bot}}(x)\Big\},\ \Gamma_{\text{soil}} := \Big\{(x,z) \in \Omega \mid z = h_{\text{soil}}(x)\Big\},\ \Gamma_{\text{ver}} := \Big\{(x,z) \in \Omega \mid x \in \partial\Omega_x\Big\}.$$

Our model split the description of the flow in two subregions of Ω (possibly time-dependent) in each of which the flow presents different behavior. We denote by h the depth of the free interface separating the freshwater layer and the unsaturated part of the aquifer. The definition of these zones is thus based on the function $h = h(t,x)$ which is an unknown of our problem. We then introduce, for a given function $h = h(t,x)$ such that $h_{\text{bot}} \leq h \leq h_{\text{soil}}$:

$$\Omega_t^- := \big\{(x,z) \in \Omega \mid z < h(x,t)\big\} \quad \text{and} \quad \Omega_t := \big\{(x,z) \in \Omega \mid z > h(x,t)\big\}, \tag{1.3}$$

and

$$\Gamma_t := \big\{(x,z) \in \Omega \mid z = h(x,t)\big\}. \tag{1.4}$$

Permeability tensor K_0. The nonlinear hydraulic conductivity is given by $\kappa K = \frac{\kappa \rho g}{\mu} K_0$. The soil transmission properties are characterized by the porosity function ϕ and the permeability tensor $K_0(x,z)$. The matrix K_0 is a 3×3 symmetric positive definite tensor which describes the conductivity of the *saturated* soil at the position $(x,z) \in \Omega$. We introduce $K_{xx} \in \mathscr{M}_{22}(\mathbb{R})$, $K_{zz} \in \mathbb{R}^*$ and $\boldsymbol{K}_{xz} \in \mathscr{M}_{21}(\mathbb{R})$ such that

$$\boldsymbol{K}_0 = \begin{pmatrix} \boldsymbol{K}_{xx} & \boldsymbol{K}_{xz} \\ \boldsymbol{K}_{xz}^T & K_{zz} \end{pmatrix}. \tag{1.5}$$

Richards hypothesis. The Richards model is moreover based on the assumption that the air pressure in the underground equals the atmospheric pressure, thus is not an unknown of the problem. One thus assumes that the moisture content and the relative conductivity of the soil are given as *functions* of the fluid pressure P, denoted respectively by $\theta = \theta(P)$ and $\kappa = \kappa(P)$. We introduce the saturation pressure P_s which is a fixed real number. The fully-saturated part of the medium corresponds to the region $\{\boldsymbol{x},\ P(\cdot,\boldsymbol{x}) > P_s\}$, while it is partially-saturated in the capillary fringe $\{\boldsymbol{x},\ P_d < P(\cdot,\boldsymbol{x}) \leq P_s\}$. The dry part is defined by the set $\{\boldsymbol{x},\ P(\cdot,\boldsymbol{x}) \leq P_d\}$. The moisture content is such that

$$\theta = \begin{cases} \phi & \text{(saturated zone)} & \text{if} \quad P(\cdot,\boldsymbol{x}) > P_s, \\ \theta(P) & \text{(with } \theta_0 \leq \theta(P) \leq \phi \text{ and } \theta'(P) > 0\text{)} & \text{if} \quad P_d < P(\cdot,\boldsymbol{x}) \leq P_s, \\ \theta_0 & \text{(dry zone)} & \text{if} \quad P(\cdot,\boldsymbol{x}) \leq P_d, \end{cases} \tag{1.6}$$

where $\theta_0 > 0$ corresponds to a residual moisture content which is positive. The associated relative hydraulic mobility is then defined by

$$\kappa(P) = \begin{cases} 1 & \text{(saturated zone)} & \text{if} \quad P(\cdot,\boldsymbol{x}) > P_s, \\ \kappa(P) & \text{(with } 0 \leq \kappa(P) \leq 1 \text{ and } \kappa'(P) > 0\text{)} & \text{if} \quad P_d < P(\cdot,\boldsymbol{x}) \leq P_s, \\ 0 & \text{(dry zone)} & \text{if} \quad P(\cdot,\boldsymbol{x}) \leq P_d. \end{cases} \tag{1.7}$$

Soil Compressibility. We neglect in the model the effects of the rock compressibility, the porosity of the medium ϕ do not depend on the pressure variations and it is thus assumed to be a constant.

Boundary conditions. To take into account the flow comming from or going to the overland, we consider a general Robin condition on the boundary Γ_{soil}:

$$a\,P + q \cdot \vec{\nu} = F \quad \text{for } (t,x,z) \in (0,T) \times \Gamma_{\text{soil}}, \tag{1.8}$$

with $a \in \mathbb{R}$ and F a source term. In the other hand, an impermeable layer is considered on the bottom of the aquifer Γ_{bot}. To simplify the presentation, we also consiter such a layer on the lateral part Γ_{ver}:

$$q \cdot \vec{\nu} = 0 \quad \text{for } (t,x,z) \in (0,T) \times \Gamma_{\text{bot}} \cup \Gamma_{\text{ver}}. \tag{1.9}$$

3d Richars problem. Before giving the model being the center of interest of this article, we recall the 3d-Richards equations which are classically used to described the water flow in an aquifer.

$$\begin{cases} \partial_t \theta(P) + \nabla \cdot \boldsymbol{v} = 0 & \text{in } (0,T)\times\Omega \\ \boldsymbol{v} = -\kappa(P)\,\boldsymbol{K}_0\Big(\dfrac{1}{\rho g}\nabla P + \boldsymbol{e}_3\Big) & \text{in } (0,T)\times\Omega \\ \alpha P + \beta\,\boldsymbol{v}\cdot\boldsymbol{n} = F & \text{on } (0,T)\times\Gamma_{\text{soil}} \\ \boldsymbol{v}\cdot\boldsymbol{n} = 0 & \text{on } (0,T)\times(\Gamma_{\text{bot}}\cup\Gamma_{\text{ver}}) \\ P(0,x,z) = P_{\text{init}}(x,z) & \text{for } (x,z)\in\Omega \end{cases} \tag{1.10}$$

3d Richars problem for weakly compressible fluid. We end this section by recalling a variant of the 3d-Richards equations where the fluid is assumed to be weakly compressible. The parameter α_p characterize this compressibility and we have

$$\begin{cases} \partial_t \theta(P) + \theta\alpha_P\,\partial_t P + \nabla\cdot\boldsymbol{v} = 0 & \text{in } (0,T)\times\Omega, \\ \boldsymbol{v} = -\kappa(P)\,\boldsymbol{K}_0\nabla\Big(\dfrac{P}{\rho g} + z\Big) & \text{in } (0,T)\times\Omega, \\ P(0,x,z) = P_{\text{init}}(x,z) & \text{for } (x,z)\in\Omega. \\ + \text{boundary conditions(1.8)and(1.9)}. \end{cases} \tag{1.11}$$

A justification of this model can be found in [17]. We also recall it in Subsection A.1.

1.2. A model coupling fast and slow component of the flow in shallow aquifers

Averaged conductivity. We introduce

$$\boldsymbol{S}_0 = \boldsymbol{K}_{xx} - \frac{1}{K_{zz}}\boldsymbol{K}_{xz}(\boldsymbol{K}_{xz}^T) \quad \text{and} \quad \boldsymbol{M}_0 = \begin{pmatrix} \boldsymbol{S}_0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{1.12}$$

The 2×2 tensor $\boldsymbol{S}_0$ is the Schur complement of the block K_{zz} in the tensor $\boldsymbol{K}_0$. It will act as an effective permeability tensor. We also introduce the averaged conductivity tensor $\tilde{K}$ defined in $(0,T)\times\Omega_x$ for any function $\tilde{H} = \tilde{H}(t,x)$ by

$$\tilde{\boldsymbol{K}}(\tilde{H})(t,x) = \int_{h_{\text{bot}}(x)}^{h_{\text{soil}}(x)} \kappa\big(\rho\, g(\tilde{H}(t,x)-z)\big)\,\boldsymbol{S}_0(x,z)\,dz. \tag{1.13}$$

- In Ω_t, the following 1d-Richards equation holds

$$\begin{cases} \partial_t\theta(P) + \partial_z\big(\boldsymbol{q}\cdot e_3\big) = 0 & \text{in } (0,T)\times\Omega_t, \\ a\,P + \boldsymbol{q}\cdot\boldsymbol{e}_3 = F & \text{in } (0,T)\times\Gamma_{\text{soil}}, \\ P\big(t,x,h(t,x)\big) = \rho\, g(\tilde{H}-h) & \text{in } (0,T)\times\Omega_x, \\ P(0,x,z) = P_{\text{init}}(x,z) & \text{in } \Omega_0. \end{cases} \tag{1.14}$$

- In Ω_t^-, the water pressure P satisfies

$$P(t,x,z) = \rho\, g\big(\tilde{H}(t,x)-z\big) \qquad \text{for } t\in[0,T[\,, \quad (x,z)\in\Omega_t^-. \tag{1.15}$$

- The hydraulic head $\tilde{H}$ is a solution in Ω_x to the following problem:

$$\begin{cases} -\nabla'\cdot(\tilde{\boldsymbol{K}}\nabla'\tilde{H}) = -(\boldsymbol{q}\cdot\boldsymbol{e}_3)|_h^+ & \text{for } (t,x)\in]0,T[\times\Omega_x, \\ \tilde{\boldsymbol{K}}(\tilde{H})\,\nabla'\tilde{H}\cdot\vec{\nu} = 0 & \text{for } (t,x)\in]0,T[\times\partial\Omega_x \\ \tilde{H}(0,x) = H_{\text{init}}(x) & \text{for } x\in\Omega_x \end{cases} \tag{1.16}$$

  where $(\boldsymbol{q}\cdot e_3)|_{\Gamma_h^+}$ denotes the trace of $\boldsymbol{q}\cdot e_3$ on Γ_t from above.
- The level $z = h$ below which we consider the vertical flow to be instantaneous is set so that

$$h(t,x) = \max\Big\{\min\Big\{\tilde{H}(t,x) - \frac{P_s}{\rho\, g}, h_{soil}(x)\Big\}, h_{\text{bot}}(x)\Big\}. \tag{1.17}$$

- The water velocity v is defined in Ω by

$$\boldsymbol{v} = \boldsymbol{q} + \boldsymbol{w} \qquad \text{for } t\in]0,T[\,, (x,z)\in\Omega, \tag{1.18}$$

  and for $t\in]0,T[$, $(x,z)\in\Omega$, the auxiliary velocities are given by

$$\boldsymbol{q} = -\kappa(P)\,K_{zz}\Big(\frac{1}{\rho\, g}\partial_z P + 1\Big)\boldsymbol{e}_3, \quad \boldsymbol{w} = -\kappa\big(\rho\, g(\tilde{H}-z)\big)\,\boldsymbol{M}_0\nabla'\tilde{H}. \tag{1.19}$$

We continue by giving a brief overview of the properties of the solution of problem (1.14)–(1.19). We refer to [7] for more details. This model is an alternative to the 3D-Richards problem for describing the flow in a shallow aquifer in a large range of time scales. The main interest being that this model is simpler to handle numerically than the 3D-Richards model. Since (1.14)–(1.19) is the coupling of a 2d problem with a lot of 1d vertical Richards problem, an important time saving in the numerical computation is expected. In the other hand, this model behaves like the 3D-Richards model for any time scale when the *ratio* ε of the deepness over the horizontal length of the aquifer is small. More precisely, it is showed in [7] that the 3D-Richards problem and this coupled model admit exactly the same effective problems when $\varepsilon \to 0$ whatever the considered time scale. Those effective problem are recalled in the Annex section (see (A.57)–(A.62)).

The model (1.14)–(1.19) is a model in the physical variables (it is not an effective model) that couple the too kind of flows that appear in the effective models in the short an long time scale (see (A.57)–(A.62)). The first one is a vertical 1d-Richards problem in the upper part of the aquifer and is associated to the velocity $\boldsymbol{q}$. It mimics the behavior of the short time scale case. The second one is a 2d-horizontal problem that assume an instantaneous vertical flow in the lower part of the aquifer. The associated velocity is $\boldsymbol{w}$ and mimics the behavior of the long time scale case.

1.3. Difficulties for the theoretical study of such a model

There is several difficulties linked to this system, the main are the following one

- the noncylindrical domain
- the time degeneracies
- the lack of control in the horizontal direction

All these difficulties were already present in the study of the paper [17] except for the time degeneration induced by taking into account the low compressibility of water in the equation in h. In the paper [17], the vertical averaging of the hydraulic conductivity made it possible to neutralize this degeneration. Here we consider a very general form for hydraulic conductivity (as done in [7]) which induces an additional nonlinearity in the time derivative term. Let us also note that, contrary to the article [17], we consider very general Robin boundary condition at the interface Γ_{soil} thus making it possible to envisage exchanges between surface water and groundwater.
The analysis of Richards equations is known to be delicate in particular due to the degenerate in time term. Usually the Kirchov transform is used to eliminate the nonlinearity in the diffusive term.
Moreover in our case, the integration domain depends on the time. There is several methods to abord the general studies of problems with free boundary. We choose here the framework of non cylindrical domain introduced by Lions and Mignot.
Finally, the first equation (1.14) is in some sense ill-posed from a theoritical point of view. Indeed since the horizontal hydraulic conductivity is null, it is an 1D-equation but defined in a 3D domain. Hence there is a lack of control of the unknown's gradient in the horizontal space variables.

2. A new model, more adapted to theoretical study

Our goal in this paper is the prove the existence of the solution of a model of the kind of (1.13)-(1.19), introduced in [7]. As said in Subsection 1.3 this problem is difficult to study as it stands. Our strategy here is to propose a new model, physically very close to (1.13)-(1.19), but for which the theoretical study is reachable. This is the goal of this section. We present a generalization of (1.14)–(1.19) in two ways:

- we allow now a non-vanishing horizontal conductivity in the upper part of the aquifer.
- we take into account of a small compressibility of the fluid in the aquifer.

This new model is given in (2.2)–(2.8) and is the one for which the existence result is given in section 3 and proved in section 4.

2.1. The new model, allowing a non-vanishing horizontal conductivity and considering small compressibility

We present and justify here a generalization of problem (1.14)–(1.19) in the compressible framework of (1.11) and in the case where the horizontal conductivity is not assumed to be vanishing in the upper part of the aquifer Ω_t. This horizontal component of the flow in Ω_t is characterized by the 2×2 symmetric positive definite tensor

$\boldsymbol{N}_0$. We introduce also

$$\boldsymbol{A}_0 = \boldsymbol{M}_0 - \begin{pmatrix} \boldsymbol{N}_0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{S}_0 - \boldsymbol{N}_0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \boldsymbol{B}_0 = \begin{pmatrix} \boldsymbol{N}_0 & 0 \\ 0 & K_{zz} \end{pmatrix} \quad \text{and} \quad \boldsymbol{G}_0 = \begin{pmatrix} \boldsymbol{N}_0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.1}$$

We consider the fluid compressibility parameter $\alpha_P > 0$ and we assume that $\alpha_P \ll 1$. We recall that this framework is the one for which the compressible 3d-Richards equations (1.11) has been obtained (see Subsection A.1).

The generalized problem is given by equations (2.2)–(2.8) bellow.

- The water velocity $\boldsymbol{v}$ is defined in Ω by

$$\boldsymbol{v} = \boldsymbol{q} + \boldsymbol{w} \qquad \text{for } t \in]0, T[\,, (\boldsymbol{x}, z) \in \Omega, \tag{2.2}$$

  and for $t \in]0, T[$, $(\boldsymbol{x}, z) \in \Omega$, the auxiliary velocities are given by

$$\boldsymbol{q} = -\kappa(P)\, \boldsymbol{B}_0 \Big(\frac{1}{\rho\, g} \nabla P + \boldsymbol{e}_3 \Big), \quad \boldsymbol{w} = -\kappa\big(\rho\, g(\tilde{H} - z)\big)\, \boldsymbol{A}_0 \nabla \tilde{H}. \tag{2.3}$$

- In $\Omega_t(t)$, the following 3d-Richards equation holds

$$\begin{cases} \partial_t \theta(P) + \alpha_P \theta(P) \partial_t P + \nabla \cdot \boldsymbol{q} = 0 & \text{for } t \in]0, T[\,, \quad (\boldsymbol{x}, z) \in \Omega_t(t) \\ \alpha\, P + \beta\, \boldsymbol{q} \cdot \boldsymbol{n} = F & \text{for } (t, \boldsymbol{x}, z) \in]0, T[\times \Gamma_{\text{soil}} \\ \boldsymbol{q} \cdot \boldsymbol{n} = 0 & \text{for } (t, \boldsymbol{x}) \in]0, T[\times \Gamma_{\text{ver}} \\ P\big(t, \boldsymbol{x}, h(t, \boldsymbol{x})\big) = \rho\, g\big(\tilde{H}(t, \boldsymbol{x}) - h(t, \boldsymbol{x})\big) & \text{for } (t, \boldsymbol{x}) \in]0, T[\times \Omega_x \\ P(0, \boldsymbol{x}, z) = P_{\text{init}}(\boldsymbol{x}, z) & \text{for } (\boldsymbol{x}, z) \in \Omega_t(0) \end{cases} \tag{2.4}$$

- In $\Omega_t^-(t)$, the water pressure P satisfies

$$P(t, \boldsymbol{x}, z) = \rho\, g\big(\tilde{H}(t, \boldsymbol{x}) - z\big) \qquad \text{for } t \in [0, T[\,, \quad (\boldsymbol{x}, z) \in \Omega_t^-(t) \tag{2.5}$$

- The hydraulic head $\tilde{H}$ is a solution in Ω_x to the following problem:

$$\begin{cases} \rho g \alpha_P (h - h_{\text{bot}}) \partial_t \tilde{H} - \nabla' \cdot \Big(\tilde{\boldsymbol{J}}(\tilde{H}) \nabla' \tilde{H} \Big) = -(\boldsymbol{q} \cdot (\boldsymbol{e}_3 - \nabla' h))|_{\Gamma_h^+} & \text{for } (t, \boldsymbol{x}) \in]0, T[\times \Omega_x \\ \tilde{\boldsymbol{J}}(\tilde{H}) \nabla' \tilde{H} \cdot \boldsymbol{n} = 0 & \text{for } (t, \boldsymbol{x}) \in]0, T[\times \partial\Omega_x \\ \tilde{H}(0, \boldsymbol{x}) = H_{\text{init}}(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in \Omega_x \end{cases} \tag{2.6}$$

  where $\big(\boldsymbol{q} \cdot (\boldsymbol{e}_3 - \nabla' h)\big)|_{\Gamma_h^+}$ denotes the normal trace of $\boldsymbol{q}$ on Γ_t from above.
- The averaged conductivity is

$$\tilde{\boldsymbol{J}}(\tilde{H})(t, \boldsymbol{x}) = \tilde{\boldsymbol{K}}(\tilde{H}) - \int_{h(t,\boldsymbol{x})}^{h_{\text{soil}}(\boldsymbol{x})} \kappa\big(\rho\, g(\tilde{H}(t, \boldsymbol{x}) - z)\big)\, \boldsymbol{N}_0(\boldsymbol{x}, z)\, dz. \tag{2.7}$$

- The level $z = h$ below which we consider the vertical flow to be instantaneous is set so that

$$h(t, \boldsymbol{x}) = \max\Big\{ \min\Big\{ \tilde{H}(t, \boldsymbol{x}) - \frac{P_s}{\rho\, g}, h_{soil}(\boldsymbol{x}) \Big\}, h_{\text{bot}}(\boldsymbol{x}) \Big\}. \tag{2.8}$$

We assume that $\boldsymbol{N}_0$ is small enough to have that $\tilde{\boldsymbol{J}}$ and $\boldsymbol{B}_0$ are positive definite.

2.2. Comments on the new model

Let us give some roughly comments about the above problem (2.2)–(2.8), in particular to highlight its differences with the original problem (1.14)–(1.19).

- The first difference concerns the compressibility. This one appears in the first equation of (2.4). This means that we consider a compressible 3d Richards problem in the upper part of the aquifer. The compressibility impact also the behavior in the lower part of the aquifer through the frist term of the first equation of (2.6). In particular problem (2.2)–(2.8), as was (1.14)–(1.19), is mass-conservative.
- The velocity vector $\boldsymbol{v}$ is still the superposition of the two velocities $\boldsymbol{q}$ and $\boldsymbol{w}$ (see equations (1.18) and (2.2)).
- Unlike in (1.14)–(1.19), the component $\boldsymbol{q}$ does not mimic exactly the behavior of the 3d-Richards model in the short-time scale, i.e. a 1d-Richards vertical problem. Indeed, the velocity $\boldsymbol{q}$ contains a non-vanishing horizontal component coming from the tensor $\boldsymbol{G}_0$. This means that we allow now a small horizontal flow in the upper part of the aquifer. The idea is that the horizontal components of the flow are non dominant in shallow aquifers in the short-time scale. This adding of $\boldsymbol{G}_0$ is possible in the equation without any impact on the effective problem in the short time scale case. In the other hand the

hortizontal components are not negligible in the long-time scale case. This introduction of $\boldsymbol{G}_0$ has an impact on the effective problem in the long time scale case.
- As in (1.14)–(1.19), the component $\boldsymbol{w}$ is horizontal. Nevertheless, it is characterized by the tensor $\boldsymbol{A}_0$ instead of $\boldsymbol{M}_0$. Due to $\boldsymbol{G}_0$, this component of the velocity is no more the only horizontal one that appear in the problem. This justify the definition (2.1) of $\boldsymbol{A}_0$.
- The same kind of difference holds for the averaged conductivity $\tilde{J}$. Indeed, it takes into account the horizontal component in $\boldsymbol{q}$ that appears in Ω_t.

The precise justification of this model is postponed to the Annex section. The strategy is the same than that of [7]. It consists in computing the effective problems associated to both compressible 3d-Richards problems and system (2.2)–(2.8) when the ratio *deepness/horizontal length* of the aquifer is very small; and for different time scale (short, intermediate and large). We then compare those effective problems and conclude that the problem (2.2)–(2.8) is a good approximation of 3d-Richards in shallow aquifers (in a large range of time scale).

At this stage we have,

- whatever[*] the choice of $\boldsymbol{G}_0$, the problem (2.2)–(2.8) is a good approximation of the compressible 3d-Richards problem (1.11) for describing the flow in shallow aquifers:
  - in the short time scale if $\alpha_p = 0$
  - in the intermediate and the long time scale, whatever the choice of α_p
- the problem (2.2)–(2.8) has the same structure than the original one (1.14)–(1.19). It also can be seen as a small perturbation of (1.14)–(1.19) when we consider a small tensor $\boldsymbol{G}_0$ and a small compressibility parameter α_P.
- by contruction, the equation in the upper part of the aquifer is now a compressible 3d-Richards problem. This increases the a priori controls on the solution[†] and reduce the degeneracy[‡]. This make the problem (2.2)–(2.8) less difficult to study theoretically than (1.14)–(1.19) (see Subsection 1.3).

Although this model (2.2)–(2.8) is physically very close to (1.14)–(1.19) and it has the same structure. Moreover it is "better posed" than (1.14)–(1.19) from a mathematical point of view.

3. Mathematical setting and main results

The main result of the paper is given in Theorem 3.7 and concern the existence of the solution of the problem (2.2)–(2.8) under some additionnal assumption. The next Subsection aim to give those assumptions and to derive a reformulation of the problem.

In this Section and in Section 4 we lighten the ratings by removing the underscript 0 on the conductivity tensors and we also will not use boldface characters to represent vectors and matrices. In particular

$$B = \boldsymbol{B}_0, \quad S = \boldsymbol{S}_0, \quad \tilde{J} = \tilde{\boldsymbol{J}}.$$

3.1. Assumptions and reformulation.

We assume that the level h characterising the level under which the vertical flow is assumed to be instantaneous, stay far from h_{bot}. We also assume that this level do not reach h_{soil}. More precisely we assume that there exists $\delta > 0$ such that

$$h_{\text{bot}} + \delta \le h(t,x) < h_{soil} \quad \forall (t,x) \in (0,T) \times \Omega.$$

Then by (2.8) it comes $h = \tilde{H} - \frac{P_s}{\rho g}$.

Taking into account of this hypothesis, the final model ($\mathscr{M}$) coupling compressible 3d-Richards flow and Dupuit horizontal flow consists in system (3.1), and (3.4) bellow

- In Ω_t the following 3d-Richards equation holds

$$\begin{cases} \partial_t \theta(P) + \theta \alpha_P \partial_t P + \nabla \cdot q = 0 & \text{in} \quad (0,T) \times \Omega_t, \\ aP + q \cdot \vec{\nu} = F & \text{on} \quad (0,T) \times \Gamma_{\text{soil}}, \\ q \cdot \vec{\nu} = 0 & \text{on} \quad (0,T) \times \Gamma_{\text{ver}}, \\ P\big(t,x,h(t,x)\big) = P_s & \text{in} \quad (0,T) \times \Omega_x, \\ P(0,x,z) = P_0(x,z) & \text{in} \quad \Omega_0. \end{cases} \tag{3.1}$$

[*] small enough to have $\tilde{\boldsymbol{J}}$ and $\boldsymbol{A}_0$ to be positive definite tensors
[†] due to non-vanishing horizontal conductivity
[‡] thanks to the non-vanishing compressibility

The effective velocity q is given by

$$q = -\kappa(P)B\nabla\left(\frac{P}{\rho g} + z\right). \tag{3.2}$$

- In Ω_t^- the pressure P satisfies

$$P(t,x,z) = \rho g\left(\frac{P_s}{\rho g} + h - z\right) \quad \text{in} \quad (0,T)\times\Omega_t^-. \tag{3.3}$$

- The depth of Γ_t, h, satisfies in Ω_x

$$\begin{cases} S_0 B_f \partial_t h - \nabla'\cdot(\tilde{J}\nabla' h) = & -\displaystyle\int_{h(t,x)}^{h_{\text{soil}}(x)} \left(\partial_t\theta(P) + \theta(P)\alpha_P\partial_t P\right) dz - q_{|z=h^+_{\text{soil}}}\cdot\vec{\nu} - \nabla'\cdot\left(\int_h^{h_{\text{soil}}} q\, dz\right), \\ \tilde{J}\nabla' h\cdot\vec{\nu} = 0 & \text{on} \quad (0,T)\times\partial\Omega_x, \\ h(0,x) = h_0(x) & \text{in} \quad \Omega_x. \end{cases} \tag{3.4}$$

The problem (3.1)–(3.4) being a problem with free boundary, we are going to define the general framework of parabolic equation in noncylindrical domain, introduced by Lions and Mignot respectively in [14] and [16].

3.2. Notations and auxiliary results

Let $\mathcal{O}$ be the open domain of $\mathbb{R}\times\mathbb{R}^N$, included in the set $\mathbb{R}^+\times\Omega$ defined by

$$\mathcal{O} = \mathbb{R}^+\times\Omega_x\times(h, h_{\text{soil}}),$$

where h is the position of the interface Γ_t. We set

$$\begin{aligned}
&\Omega_{t'} = \mathcal{O}\cap\{t = t'\},\ \forall t'\geq 0, \ \text{(definition compatible with (1.3))}\\
&\Omega'_{t'} = \Omega\setminus\Omega_{t'},\ \forall t'\geq 0,\\
&\mathcal{O}_T = \mathcal{O}\cap\{0\leq t\leq T\},\\
&\mathcal{O}'_T = \big((0,T)\times\Omega\big)\setminus\mathcal{O}_T,\\
&\Gamma' = \Gamma\setminus\Omega_0 \ \text{(i.e. the lateral boundary of}\ \mathcal{O}),\\
&\gamma_t = \partial\Omega_t \ \text{(i.e. the boundary of}\ \Omega_t\subset\mathbb{R}^N),\\
&\Gamma'_T = \Gamma'\cap\{0<t<T\} = \cup_{t\in(0,T)}\gamma_t.
\end{aligned}$$

We define

$$H^{0,1}(\mathcal{O}) = \{u\,|\,D^p u\in L^2(\mathcal{O}) \text{ for } |p|\leq 1\},$$

where

$$D^p u = \{D^\alpha u | \alpha = (\alpha_1,\alpha_2,\alpha_2) \text{with } |\alpha| = p\}.$$

It is an Hilbert space endowed with the norm

$$\|u\|_{H^{0,1}(\mathcal{O})} = \big(\Sigma_{p\leq 1}\int_{\mathcal{O}} |D^p u| dx dt\big)^{1/2}.$$

$H_0^{0,1}(\mathcal{O})$ denotes the closure of $\mathcal{D}(\mathcal{O})$ in $H^{0,1}(\mathcal{O})$ for the norm $\|\ \|_{H^{0,1}(\mathcal{O})}$. In the following, we will denote $F(\mathcal{O}) = H_0^{0,1}(\mathcal{O})$ and $F'(\mathcal{O})$ its topological dual. Besides, we introduce

$$\mathcal{A}(\mathcal{O}) = \{u|\, u\in H^{0,1}(\mathcal{O}),\ \frac{du}{dt}\in F'(\mathcal{O})\},$$
$$\mathcal{B}(\mathcal{O}) = \{u|\, u\in F(\mathcal{O}),\ \frac{du}{dt}\in F'(\mathcal{O})\},$$

endowed with the Hilbertian norms
$\|\cdot\|_{\mathcal{A}(\mathcal{O})} = \big(\|\cdot\|^2_{H^{0,1}(\mathcal{O})} + \|\partial_t\cdot\|^2_{(H^{0,1}(\mathcal{O}))'}\big)^{1/2}$ and $\|\cdot\|_{\mathcal{B}(\mathcal{O})} = \big(\|\cdot\|^2_{H_0^{0,1}(\mathcal{O})} + \|\partial_t\cdot\|^2_{F'(\mathcal{O})}\big)^{1/2}$.
Finally, $B_0(\mathcal{O})$ (resp. $B_T(\mathcal{O})$) is the closure in $\mathcal{B}(\mathcal{O})$ of functions null in a neighborhood of $t=0$ (resp. t=T).
We now state some auxiliary results proved in [14]

Lemma 3.1. *If $\mathcal{O}$ is sufficiently regular, we thus have*

- *1.* $H^{0,1}(\mathscr{O}_T) = L^2([0,T];H^1(\Omega_t))$ *where*
$$L^2([0,T];H^1(\Omega_t)) = \{u \,|\, u(t,.) \in H^1(\Omega_t),\ t \in [0,T] \ \textit{a.e. and}\ \|u\|_{H^{0,1}(\mathscr{O}_T)} < +\infty\},$$
with $\|u\|_{H^{0,1}(\mathscr{O}_T)} = \int_0^T \|u\|^2_{H^1(\Omega_t)}\, dt$.
A similar result holds for $H_0^{0,1}(\mathscr{O}_T)$.
- *2. For* $u \in H_0^{0,1}(\mathscr{O})$, *we can define* $\gamma(u)$, *the trace of* u *on* Γ' *in* $L^2(\Gamma')$.
Moreover $u \in F(\mathscr{O}) \Longleftrightarrow \gamma(u) = 0$.
- *3. Let* $u \in \mathscr{B}(\mathscr{O}_T)$, *thus* $u \in B_T(\mathscr{O}) \Longleftrightarrow u(T,.) = 0$.
- *4.* $\forall\, u, v \in \mathscr{B}(\mathscr{O}_s)$, *we have*
$$\langle \frac{\partial u}{\partial t}, v\rangle_{F',F} + \langle u, \frac{\partial v}{\partial t}\rangle_{F',F} = (u(s,.), v(s,.))_{L^2(\Omega_s)} - (u(0,.), v(0,.))_{L^2(\Omega_0)}. \tag{3.5}$$

Let Ω' an open bounded domain of $\mathbb{R}^3$. For the sake of brevity we shall write $H^1(\Omega') = W^{1,2}(\Omega')$ and
$$V(\Omega') = H_0^1(\Omega'),\ V'(\Omega') = H^{-1}(\Omega'),\ H(\Omega') = L^2(\Omega').$$
The embeddings $V(\Omega') \subset H(\Omega') = H'(\Omega') \subset V'(\Omega')$ are dense and compact. For any $T > 0$, let $W(0,T,\Omega')$ denote the space
$$W(0,T,\Omega') := \{\omega \in L^2(0,T;V(\Omega')),\ \partial_t \omega \in L^2(0,T;V'(\Omega'))\}$$
endowed with the Hilbertian norm $\|\cdot\|_{W(0,T,\Omega')} = \left(\|\cdot\|^2_{L^2(0,T;V(\Omega'))} + \|\partial_t \cdot\|^2_{L^2(0,T;V'(\Omega'))}\right)^{1/2}$. The following embeddings are continuous ([15] prop. 2.1 and thm 3.1, chapter 1)
$$W(0,T,\Omega') \subset \mathscr{C}([0,T];[V(\Omega'),V'(\Omega')]_{\frac{1}{2}}) = \mathscr{C}([0,T];H(\Omega'))$$
while the embedding
$$W(0,T,\Omega') \subset L^2(0,T;H(\Omega')) \tag{3.6}$$
is compact (Aubin's Lemma, see [20]).
The following result by F. Mignot (see [10]) is used in the sequel.

Lemma 3.2. *Let* $f : R \to R$ *be a continuous and nondecreasing function such that* $\limsup_{|\lambda|\to+\infty} |f(\lambda)/\lambda| < +\infty$.
Let $\omega \in L^2(0,T;H(\Omega'))$ *be such that* $\partial_t \omega \in L^2(0,T;V'(\Omega'))$ *and* $f(\omega) \in L^2(0,T;V(\Omega'))$.
Then
$$\langle \partial_t \omega, f(\omega)\rangle_{V(\Omega')',V(\Omega')} = \frac{d}{dt}\int_\Omega \Big(\int_0^{\omega(\cdot,y)} f(r)\, dr\Big)\, dy \ \textit{in}\ \mathscr{D}'(0,T).$$
Hence for all $0 \le t_1 < t_2 \le T$
$$\int_{t_1}^{t_2} \langle \partial_t \omega, f(\omega)\rangle_{V',V}\, dt = \int_\Omega \Big(\int_{\omega(t_1,y)}^{\omega(t_2,y)} f(r)\, dr\Big)\, dy.$$

Remark 3.3. *The result* (3.5) *of Lemma 3.1 is a generalization of Lemma 3.2 to the case where the space domain* Ω' *is time dependent.*

3.3. Main results

We aim giving an existence result of physically admissible weak solutions for model $(\mathscr{M})$ completed by initial and boundary conditions.
Let us first detail the mathematical assumptions. We begin with the characteristics of the porous structure. We first assume that depths h_{bot} and h_{soil} are constant and such that $h_{\text{bot}} > h_{\text{soil}} > 0$. We recall that
$$q = -\kappa(P)\, B \nabla\Big(\frac{P}{\rho\, g} + z\Big).$$
We recall that the 2×2 averaged conductivity tensor $\tilde{J}$ defined in $(0,T) \times \Omega_x$ for any function $\tilde{H} = \tilde{H}(t,x)$ is given by
$$\tilde{J}(\tilde{H})(t,x) = \int_{h_{\text{bot}}}^{h(x)} S(x,z)\, dz + \int_{h(x)}^{h_{\text{soil}}} \kappa\big(\rho\, g(\tilde{H}(t,x) - z)\big)\,(S(x,z) - N)\, dz. \tag{3.7}$$
Furthermore functions θ and κ are pressure-dependent and we assume that
$$\theta \in \mathscr{C}^1(\mathbb{R}), \quad 0 < \theta_- := \phi s_0 \le \theta(x) \le \theta_+, \quad \theta'(x) \ge 0\ \forall x \in \mathbb{R}, \tag{3.8}$$
$$\kappa \in \mathscr{C}(\mathbb{R}), \quad 0 < \kappa_- \le \kappa(x) \le \kappa_+\ \forall x \in \mathbb{R}. \tag{3.9}$$

Before stating the main result of this work, we will transform the original problem and bring us back to the framework introduced in [16].
The above assumptions on the fluid and the medium allow to eliminate the nonlinearity in time of Eq. (3.1), namely Assumptions (3.8)-(3.9) are sufficient to define the primitive function $\mathscr{P}$ such that

$$\mathscr{P}(P) = \theta(P) + \alpha_P \int^P \theta(s)\, ds.$$

A direct computation gives $\mathscr{P}'(P) = \theta'(P) + \alpha_P \theta(P) > \alpha_P \theta_- > 0$, indeed by previous hypothesis, we have $\theta'(P) \geq 0$ and $\theta(P) > \phi s_0$.
Since $\mathscr{P}$ is a bijective application, the existence of p such that

$$p = \mathscr{P}(P)$$

is equivalent to the existence of P solution of the original Richards problem. The transform $\mathscr{P}$ of Eq. (3.1) is

$$\partial_t p - \frac{1}{\rho g} \nabla \cdot \left(\frac{\kappa(\mathscr{P}^{-1}(p))}{(\theta' + \alpha_P \theta)(\mathscr{P}^{-1}(p))} B \nabla p \right) - \nabla \cdot \left(\kappa(\mathscr{P}^{-1}(p)) B \vec{e}_3 \right) = 0.$$

We introduce the notation

$$\tau(p) = \frac{1}{\rho g} \frac{\kappa(\mathscr{P}^{-1}(p))}{(\theta' + \alpha_P \theta)(\mathscr{P}^{-1}(p))}.$$

Note that, due to hypotheses (3.8)-(3.9), there exist two positive reals τ_- and τ_+ such that

$$0 < \tau_- := \frac{\kappa_-}{\rho g\, \alpha_P \theta_+} \leq \tau(p) \leq \tau_+ := \frac{\kappa_+}{\rho g\, \alpha_P \theta_-}. \tag{3.10}$$

Besides, tensors B and $\tilde{J}$ are assumed to be bounded and uniformly elliptic. More precisely, there exist two couple of positive real numbers, $0 < K^- \leq K^+$ and $0 < \tilde{K}^- \leq \tilde{K}^+$ such that

$$0 < K^- |\xi|^2 \leq B\xi \cdot \xi = \sum_{k,l=1}^{3} K_{kl} \xi_k \xi_l \leq K^+ |\xi|^2, \quad \forall \xi \in \mathbb{R}^3 \setminus \{0\}, \tag{3.11}$$

and

$$0 < \tilde{K}^- |\xi|^2 \leq \tilde{J}\xi \cdot \xi = \sum_{k,l=1}^{2} \tilde{K}_{kl} \xi_k \xi_l \leq \tilde{K}^+ |\xi|^2, \quad \forall \xi \in \mathbb{R}^2 \setminus \{0\}. \tag{3.12}$$

Let $\delta \in \mathbb{R}$ be a positive number and $\ell = (h_{\text{soil}} - h_{\text{bot}})$ the parameter denoting to the total thickness of the subsoil. We introduce the function T_l defined by

$$T_l(h) = (h - h_{\text{bot}}) \quad \forall h \in [h_{\text{bot}} + \delta, h_{\text{soil}}],$$

which is extended continuously and constantly outside $[h_{\text{bot}} + \delta, h_{\text{soil}}]$. The hypothesis on the parameter δ is sufficient to define the primitive function $\mathscr{T}$ such that

$$\mathscr{T}(h) = \begin{cases} \delta(h - h_{\text{bot}}) - \dfrac{\delta^2}{2} & \text{if} \quad h_{\text{bot}} + \frac{\delta}{2} \leq h \leq h_{\text{bot}} + \delta, \\ \dfrac{(h - h_{\text{bot}})^2}{2} & \text{if} \quad h_{\text{bot}} + \delta \leq h \leq h_{\text{soil}}, \\ \ell(h - h_{\text{bot}}) - \dfrac{\ell^2}{2} & \text{if} \quad h \geq h_{\text{soil}}. \end{cases}$$

A direct computation gives $\mathscr{T}'(h) = T_l(h) > \delta > 0$. Since $\mathscr{T}$ is a bijective application, the existence of u such that $u = \mathscr{T}(h)$ is equivalent to the existence of h solution of the original Eq. (3.4), moreover we have

$$\mathscr{T}^{-1}(u) = \begin{cases} \dfrac{u}{\delta} + \dfrac{\delta}{2} + h_{\text{bot}} & \text{if} \quad 0 \leq u \leq \dfrac{\delta^2}{2}, \\ \sqrt{2u} & \text{if} \quad \dfrac{\delta^2}{2} \leq u \leq \dfrac{\ell^2}{2}, \\ \dfrac{u}{\ell} + \dfrac{\ell}{2} + h_{\text{bot}} & \text{if} \quad u \geq \dfrac{\ell^2}{2}. \end{cases}$$

Hence, we have

$$\frac{1}{\ell} \leq \| \left(\mathscr{T}^{-1} \right)' \|_\infty \leq \frac{1}{\delta}. \tag{3.13}$$

Remark 3.4. *The small parameter δ is introduced to control the degenerate in time term in (3.4). The physical significance of this addition is that there is an amount of water at least δ thick everywhere in the aquifer. We underline that in the article [17], it was also necessary to impose this assumption. In addition, a second degeneration is also hidden in the model. Indeed, to extend the solution p outside the time dependent domain Ω_t, we must also impose on the function h to be more than or equal to a quantity strictly greater than h_{bot}. This is another reason why the small parameter δ was introduced in the lower bound of the study interval.*

Let $\chi(u)$ the function defined by

$$\chi(u) = \begin{cases} 0 & \text{if} \quad u \le 0 \\ 1 & \text{if} \quad u > 0 \end{cases}.$$

In order to guarantee the nonnegativity of u, the control $\chi(u)$ is added in front of the right hand side of (3.14) .

The transform $\mathscr{T}$ of Eq. (2.6) is

$$S_0 \partial_t u - \nabla' \cdot \big(\tilde{J}(\mathscr{T}^{-1}(u)) \nabla' \mathscr{T}^{-1}(u)\big) = -\chi(u)\Big(\int_{h_{\text{bot}}+T_l(\mathscr{T}^{-1}(u))}^{h_{\text{soil}}} \partial_t p \, dz \tag{3.14}$$

$$+\text{div}_x\Big(\int_{h_{\text{bot}}+T_l(\mathscr{T}^{-1}(u))}^{h_{\text{soil}}} q \, dz\Big) + q_{|z=h_{\text{soil}}^+} \cdot \vec{\nu}\Big) \tag{3.15}$$

Definition 3.5. *The definition of the depth h is derived from the construction of u. Namely, for u given by (3.14), we set*

$$h(t,x) := h_{\text{bot}} + T_l(\mathscr{T}^{-1}(u)). \tag{3.16}$$

Remark 3.6. *This definition of h allows to define the integration domain Ω_t (and then the interface Γ_t) in the system (3.17)-(3.18). We emphasize that by definition, h always remains in the interval $[h_{\text{bot}} + \delta, h_{\text{soil}}]$.*

We are led to consider the following problem completed by the boundary and initial conditions :

$$\partial_t p - \nabla \cdot \big(\tau(p) B \nabla p\big) - \nabla \cdot \Big(\kappa(\mathscr{P}^{-1}(p)) B \vec{e_3}\Big) = 0 \quad \text{in} \quad \mathscr{O}_T, \tag{3.17}$$

$$p|_{\Gamma_t} = \mathscr{P}(P_s) \quad \text{in} \quad (0,T), \qquad \nabla\big(\mathscr{P}^{-1}(p) + \rho g z\big) \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times \Gamma_{ver},$$

$$a P + \nabla\big(\mathscr{P}^{-1}(p) + b\rho_0 g z\big) \cdot \vec{\nu} = F \quad \text{on} \quad (0,T) \times \Gamma_{\text{soil}}, \quad p(0,x,z) = \mathscr{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0. \tag{3.18}$$

$$S_0 \partial_t u - \nabla' \cdot \big(\tilde{J}(\mathscr{T}^{-1}(u)) \nabla' \mathscr{T}^{-1}(u)\big) = -\chi(u)\Big(\int_{h_{\text{bot}}+T_l(\mathscr{T}^{-1}(u(t,x)))}^{h_{\text{soil}}} \partial_t p \, dz$$

$$+ \text{div}_x\Big(\int_{h_{\text{bot}}+T_l(\mathscr{T}^{-1}(u))}^{h_{\text{soil}}} q \, dz\Big) + q_{|z=h_{\text{soil}}^+} \cdot \vec{\nu}\Big) \quad in \quad (0,T) \times \Omega_x, \tag{3.19}$$

$$\nabla u \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times \partial\Omega_x, \quad u(0,x) = (h_0(x) - h_{\text{bot}}(x))^2 \quad \text{in} \quad \Omega_x, \tag{3.20}$$

where P_s is constant with respect to the time and the space. Function $P_0 \in H^2(\Omega)$ satisfies the compatibility condition

$$P_0(x, h_0) = P_s \quad \text{in} \quad \Omega_0.$$

We also assume that $h_0 \in L^\infty(\Omega_x)$ is such that

$$h_{\text{bot}} + \delta \le h_0 \le h_{\text{soil}} \text{ a.e. in } \quad \Omega_x. \tag{3.21}$$

Finally we suppose that source terms F and $P_{|\Gamma_{\text{soil}}}$ are given fonctions belonging to the space $L^2(0,T,L^2(\Omega_x))$. For the previous parabolic system, we state and prove the following existence result.

Theorem 3.7. *Assume that there exist two real numbers θ_- and κ_- such that*

$$\theta(x) \ge \theta_- > 0 \quad \forall x \in \mathbb{R}, \quad \kappa(x) \ge \kappa_- > 0 \quad \forall x \in \mathbb{R}^+. \tag{3.22}$$

Then system (3.17)- (3.18),(3.19)-(3.20) admits a weak solution (p,u) satisfying
(a) the function $p \in L^2(0,T;H^1(\Omega)) \cap L^2(0,T;(H^1(\Omega))')$ is solution of (3.17)- (3.18),
(b) the function $u \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ is solution of (3.19)- (3.20). Moreover $u(t,x) \ge 0$ a.e. in $[0,T] \times \Omega_x$.

Corollary 3.8. *Assume that there exist two real numbers θ_- and κ_- such that*

$$\theta(x) \ge \theta_- > 0 \quad \forall x \in \mathbb{R}, \quad \kappa(x) \ge \kappa_- > 0 \quad \forall x \in \mathbb{R}^+. \tag{3.23}$$

Then the model $\mathscr{M}$ admits a weak solution (P,h) such that
(a) the function $P \in L^2(0,T;H^1(\Omega)) \cap L^2(0,T;(H^1(\Omega))')$;
(b) the function $h \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ and $h(t,x) \in [h_{\text{bot}} + \delta, h_{\text{soil}}]$ a.e. in $[0,T] \times \Omega_x$.

The proof of Corollary 3.8 is a direct consequence of Theorem 3.7 since we turn back to the original problem by considering inverse transforms $\mathscr{P}^{-1}$ and $\mathscr{T}^{-1}$.

Next section is devoted to the proof of Theorem 3.7.

4. Proof of Theorem 3.7

Let us sketch the global strategy of the proof. The problem consists of a strongly nonlinear coupled system, so we apply a fixed point approach to solve it in two steps. In the first step, we decouple the system and apply a fixed point Schauder theorem to establish an existence and uniqueness result for each decoupled and regularized equation. Then we establish compactness results which allow us to prove the global existence in time of the initial problem. One of the main difficulties of the study is that we are working on time-dependent domains. This difficulty is solved by using the work of Lions and Mignot for parabolic equations on noncylindrical domains. This consists in suitably extending the solution outside the variable domain, thus bringing us back to a fixed domain ([14, 16]).

4.1. Fixed point step

We first reduce the boundary condition on interface Γ_t of the system (3.17)-(3.18) to homogeneous Dirichlet boundary condition. To do this, we set $\bar{p} = p - \mathscr{P}(P_s)$. Since $\mathscr{P}(P_s)$ is a constant, the system (3.17)-(3.18) becomes

$$\begin{aligned}
&\partial_t \bar{p} - \nabla\cdot\big(\bar{\tau}(\bar{p})\,B\nabla\bar{p}\big) - \nabla\cdot\big(\bar{\kappa}(\bar{p})B\vec{e}_3\big) = 0 \quad \text{in} \quad \mathscr{O}_T,\\
&\bar{p}|_{\Gamma_t} = 0 \quad \text{in} \quad (0,T), \qquad \nabla\big(\mathscr{P}^{-1}\big(\bar{p}+\mathscr{P}(P_s)\big) + \rho g z\big)\cdot\vec{\nu} = 0 \quad \text{on} \quad (0,T)\times\Gamma_{ver},\\
&aP + \nabla\big(\mathscr{P}^{-1}(\bar{p}+\mathscr{P}(P_s)) + b\rho g z\big)\cdot\vec{\nu} = F \quad \text{on} \quad (0,T)\times\Gamma_{\text{soil}},\\
&\bar{p}(0,x,z) = \mathscr{P}(P_0)(x,z) - \mathscr{P}(P_s) \quad \text{in} \quad \Omega_0,
\end{aligned}$$

where $\bar{\tau}(\bar{p}) = \tau\big(\bar{p}+\mathscr{P}(P_s)\big)$ and $\bar{\kappa}(\bar{p}) = \kappa o\mathscr{P}^{-1}\big(\bar{p}+\mathscr{P}(P_s)\big)$. We thus remark, that just renaming functions τ and κ, we go back to the case $\mathscr{P}(P_s) = 0$ on Γ_t. So, from now, we omit the subscript "$\bar{\cdot}$" in the previous system and we consider the original system : (3.17)-(3.18) with $\mathscr{P}(P_s) = 0$:

$$\partial_t p - \nabla\cdot\big(\tau(p)\,B\nabla p\big) - \nabla\cdot\Big(\kappa(\mathscr{P}^{-1}(p))B\vec{e}_3\Big) = 0 \quad \text{in} \quad \mathscr{O}_T, \tag{4.1}$$

$$\begin{aligned}
&p|_{\Gamma_t} = 0 \quad \text{in} \quad (0,T), \qquad \nabla\big(\mathscr{P}^{-1}(p) + \rho g z\big)\cdot\vec{\nu} = 0 \quad \text{on} \quad (0,T)\times\Gamma_{ver},\\
&aP + \nabla\big(\mathscr{P}^{-1}(p) + b\rho g z\big)\cdot\vec{\nu} = F \text{ on } (0,T)\times\Gamma_{\text{soil}},\ p(0,x,z) = \mathscr{P}(P_0)(x,z) \text{ in } \Omega_0.
\end{aligned} \tag{4.2}$$

Definition 4.1. *We call weak solution of problem (3.19)-(3.20) any solution $u \in L^2(0,T;H^1(\Omega_x)) \cap L^2(0,T;(H^1(\Omega_x))')$ that satisfies the weak formulation $\forall \phi \in L^2(0,T;H^1(\Omega_x))$*

$$\begin{aligned}
&S_0\langle\partial_t u,\phi\rangle_{(H^1)',H^1} + \int_0^T\int_{\Omega_x}\Big(\frac{\mathscr{T}^{-1}(u)}{\mathscr{T}'(\mathscr{T}^{-1}(u))}\tilde{J}\nabla' u\cdot\nabla'\phi\Big)dx\,dt\\
&= -\int_0^T\int_{\Omega_x}\chi(u)\Big(\Big(\int_{h(t,x)}^{h_{\text{soil}}}\partial_t\bar{p}\,dz + aP|_{\Gamma_{\text{soil}}} - F\Big)\phi - \Big(\int_{h(t,x)}^{h_{\text{soil}}} q\,dz\Big)\cdot\nabla'\phi\Big)dx\,dt,
\end{aligned} \tag{4.3}$$

$$u(0,x) = \frac{(h_0(x) - h_{\text{bot}}(x))^2}{2} \quad \text{in} \quad \Omega_x, \tag{4.4}$$

where $h(t,x) := h_{\text{bot}} + T_l(\mathscr{T}^{-1}(u(t,x)))$ and $q = -\kappa(P)\,B\nabla\big(\frac{P}{\rho g} + z\big)$.

We directly infer the previous weak formulation from Eq. (3.19) by keeping in mind that

$$q\cdot\vec{\nu} = 0 \quad \text{on} \quad (0,T)\times\Gamma_{\text{ver}}.$$

Definition 4.2. *We call weak solution of problem (4.1)-(4.2) any solution $p \in W(0,T,\Omega)$ s.t.*

(1) *$p = 0$ in $\Omega\setminus\Omega_t$, $\forall t \in (0,T)$,*

(2) *the solution p satisfies the weak formulation in $\mathcal{O}_T$, $\forall \phi \in \mathcal{A}(\mathcal{O})$ (null on the interface Γ_t)*

$$\langle \partial_t p, \phi \rangle_{F',F} + \int_0^T \Big(\int_{\Omega_t} \big(\tau(p) B \nabla p + \kappa(\mathscr{P}^{-1}(p)) B \vec{e}_3 \big) \cdot \nabla \phi \Big) dt$$
$$= \int_0^T \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}}) \phi_{|\Gamma_{\text{soil}}} \, dx \, dt, \tag{4.5}$$
$$p(0,x,z) = \mathscr{P}(P_0)(x,z) \text{ in } \Omega_0. \tag{4.6}$$

We now construct the framework to apply the Schauder fixed point theorem (see [9, 22]). For the fixed point strategy, we introduce two convex subsets (W_1, W_2) of $W(0,T,\Omega_x) \times W(0,T,\Omega)$, namely

$$W_1 := \{ u \in W(0,T,\Omega_x);\ u(0) = u_0, \|u\|_{L^2(0,T;H^1(\Omega_x))} \le C_u \text{ and } \|u\|_{L^2(0,T;(H^1(\Omega_x))')} \le C_u' \}. \tag{4.7}$$

and

$$W_2 := \{ p \in W(0,T,\Omega);\ p(0) = p_0, \|p\|_{L^2(0,T;H^1(\Omega))} \le C_p \quad \text{and} \quad \|p\|_{L^2(0,T;(H^1(\Omega))')} \le C_p' \}, \tag{4.8}$$

where constants (C_p, C_p') and (C_u, C_u') are defined thereafter.
Let $(\bar{u}, \bar{p}) \in W_1 \times W_2$, we begin by considering the unique solution u of the following linearized problem

$$S_0 \partial_t u - \nabla' \cdot \Big(\frac{\mathscr{T}^{-1}(\bar{u})}{\mathscr{T}'(\mathscr{T}^{-1}(\bar{u}))} \tilde{J} \nabla' u \Big)$$
$$= -\chi(u) \Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}} \partial_t \bar{p} \, dz + \text{div}_x \Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}} \bar{q} \, dz \Big) + aP|_{\Gamma_{\text{soil}}} - F \Big) \quad \text{in} \quad (0,T) \times \Omega_x, \tag{4.9}$$
$$\nabla u \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times \partial\Omega_x, \quad u(0,x) = \frac{(h_0(x) - h_{\text{bot}}(x))^2}{2} \quad \text{in} \quad \Omega_x, \tag{4.10}$$

where $\bar{h}(t,x) := h_{\text{bot}} + T_l(\mathscr{T}^{-1}(\bar{u}(t,x)))$.

Remark 4.3.

(1) *We have to precise the meaning of the term $\int_{\bar{h}(t,x)}^{h_{\text{soil}}(x)} \partial_t \bar{p} \, dz$:*

$$\int_{\bar{h}(t,x)}^{h_{\text{soil}}} \partial_t \bar{p} \, dz = \int_{h_{\text{bot}}}^{h_{\text{soil}}} \chi_{z \ge \bar{h}(t,x)} \partial_t \bar{p} \, dz$$

is the function of $(H^1(\Omega_x))'$ such that $\forall v \in H^1(\Omega_x) \subset H^1(\Omega)$

$$\langle \int_{\bar{h}(t,x)}^{h_{\text{soil}}} \partial_t \bar{p} dz, \, v \rangle_{H^1(\Omega_x)', H^1(\Omega_x)} = \langle \partial_t \bar{p}, \chi_{z \ge \bar{h}(t,x)} \, v \rangle_{H^1(\Omega)', H^1(\Omega)}.$$

(2) *In previous system (4.9)-(4.10), taking into account the Robin boundary condition on the boundary $z = h_{\text{soil}}$, we replace the flux $q_{|z=h_{\text{soil}}^+} \cdot \vec{\nu}$ by $F - aP_{|\Gamma_{\text{soil}}}$. These data express source term and exchanges between the overland water and the aquifer. Note also that since h_{soil} is assumed to be constant, the unit normal vector $\vec{\nu}$ thus corresponds to $\vec{e}_3$. In the following, we denote by $F_R = F - aP_{|\Gamma_{\text{soil}}} \in L^2(0,T,L^2(\Omega_x))$ the result of these two external inputs.*

Lemma 4.4. *Let $h_0 \in L^\infty(\Omega_x)$ satisfying (3.21), there exists a unique weak solution $u \in W(0,T,\Omega_x)$ of (4.9)-(4.10) such that*

$$\|u\|_{L^2(0,T;H^1(\Omega_x))} \le C_u \quad \text{and} \quad \|u\|_{L^2(0,T;(H^1(\Omega_x))')} \le C_u',$$

where C_u and C_u' only depend on the data of the problem.
Moreover, $u \ge 0$, a.e. in $[0,T] \times \Omega_x$.

Proof.

Step 1. Global existence and uniform estimates

It follows from the classical textbook [13] pp. 178-179 that for every nonnegative function $\bar{u} \in W(0,T,\Omega_x)$ (and $\bar{h}$ *s.t.* $\bar{h}(t,x) := h_{\text{bot}} + T_l(\mathscr{T}^{-1}(\bar{u}(t,x)))$), there exists a solution $u \in W(0,T,\Omega_x)$ of the parabolic problem with smooth coefficients

$$S_0 \partial_t u - \nabla' \cdot \Big(\frac{\mathscr{T}^{-1}(\bar{u})}{\mathscr{T}'(\mathscr{T}^{-1}(\bar{u}))} \tilde{J} \nabla' u \Big) = -\chi(u) \Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}} \partial_t \bar{p} \, dz + \text{div}_x \Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}} \bar{q} \, dz \Big) - F_R \Big) \quad \text{in} \quad (0,T) \times \Omega_x, \tag{4.11}$$
$$\nabla u \cdot \vec{\nu} = 0 \quad \text{on} \quad (0,T) \times \partial\Omega_x, \quad u(0,x) = \frac{(h_0(x) - h_{\text{bot}}(x))^2}{2} \quad \text{in} \quad \Omega_x.$$

Combining (3.12) and (3.13), we first remark that

$$\int_{\Omega_x} \frac{\mathscr{T}^{-1}(\bar{u})}{\mathscr{T}'(\mathscr{T}^{-1}(\bar{u}))} \tilde{J}\nabla' u\cdot\nabla' u \geq \frac{\tilde{K}^-}{\ell}\int_{\Omega_x}|\nabla' u|^2 dx$$

Multiplying Eq. (4.11) by u and integrating by parts over Ω_x, we obtain

$$\frac{S_0}{2}\frac{d}{dt}\int_{\Omega_x}|u(t,.)|^2\,dx+\frac{\tilde{K}^-}{\ell}\int_{\Omega_x}|\nabla' u|^2\,dx \leq |\langle\int_{\bar{h}(t,x)}^{h_{\text{soil}}}\partial_t\bar{p}\,dz,\,u\rangle_{H^1(\Omega_x)',H^1(\Omega_x)}| + |\int_{\Omega_x}\Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}}\bar{q}\,dz\Big)\cdot\nabla' u\,dx| + |\int_{\Omega_x}F_R\,u\,dx|. \tag{4.12}$$

Furthermore, by definition of the function T_l, we have

$$\begin{aligned}
&|\langle\int_{\bar{h}(t,x)}^{h_{\text{soil}}}\partial_t\bar{p}\,dz,\,u\rangle_{H^1(\Omega_x)',H^1(\Omega_x)}| \leq |h_{\text{soil}}-h_{\text{bot}}|^{1/2}\,\|u\|_{H^1(\Omega_x)}\|\partial_t\bar{p}\|_{(H^1(\Omega))'}\\
&\leq \frac{\tilde{K}^-}{4\ell}\big(\|u\|^2_{L^2(\Omega_x)}+\|\nabla' u\|^2_{L^2(\Omega_x)}\big)+\frac{\ell}{\tilde{K}^-}|h_{\text{soil}}-h_{\text{bot}}|\|\partial_t\bar{p}\|^2_{(H^1(\Omega))'},\\
&|\int_{\Omega_x}\Big(\int_{\bar{h}(t,x)}^{h_{\text{soil}}}\bar{q}\,dz\Big)\cdot\nabla' u\,dx| \leq \|\nabla' u\|_{L^2(\Omega_x)}\times\kappa_+\tilde{K}^+\Big(\frac{\|\nabla\bar{p}\|_{L^2(\Omega)}}{\rho\,g}+mes(\Omega)^{1/2}\Big)\\
&\leq \frac{\tilde{K}^-}{4\ell}\|\nabla' u\|^2_{L^2(\Omega_x)}+\frac{2\ell}{\tilde{K}^-}(\kappa_+\tilde{K}^+)^2\Big(\frac{\|\nabla\bar{p}\|^2_{L^2(\Omega)}}{\rho^2 g^2}+mes(\Omega)\Big),\\
&|\int_{\Omega_x}F_R\,u\,dx| \leq \|u\|_{L^2(\Omega_x)}\|F_R\|_{L^2(\Omega_x)} \leq \frac{1}{2}\|u\|^2_{L^2(\Omega_x)}+\frac{1}{2}\|F_R\|^2_{L^2(\Omega_x)}.
\end{aligned}$$

Applying Gronwall's inequality in its differential form, we get

$$\begin{aligned}
\|u(t,.)\|_{L^2(\Omega_x)} \leq & e^{\frac{(1+\tilde{K}^-/2\ell)\,T}{S_0}}\Big(\|u_0\|_{L^2(\Omega_x)}+\frac{1}{S_0}\|F_R\|^2_{L^2((0,T)\times\Omega_x)}+\frac{4\,\ell\,T}{S_0\tilde{K}^-}(\kappa_+\tilde{K}^+)^2 mes(\Omega)\\
&+\frac{4\,\ell}{S_0\tilde{K}^-}(\kappa_+\tilde{K}^+)^2\frac{\|\nabla\bar{p}\|^2_{L^2(\Omega_T)}}{\rho^2 g^2}+\frac{2\,\ell^2}{S_0\tilde{K}^-}\|\partial_t\bar{p}\|^2_{L^2(0,T,(H^1(\Omega))')}\Big).
\end{aligned}$$

Then from this estimate and (4.12), we deduce

$$\|\,u\|^2_{L^2(0,T;H^1(\Omega_x))} \leq C(T,S_0,u_0,\tilde{J},\ell,F_R,C_p,C_p') := C_u.$$

On the other hand

$$\begin{aligned}
&\|\frac{du}{dt}\|_{L^2(0,T;H^1(\Omega_x)')} = sup_{\|v\|_{L^2(0,T;H^1(\Omega_x))}\leq 1}|\int_0^T\langle\frac{du}{dt},v\rangle_{H^1(\Omega_x)',H^1(\Omega_x))}dt|\\
&\leq\frac{2}{S_0}\Big(\frac{\tilde{K}^+}{\delta}\|u(t,.)\|_{L^2(0,T;H^1(\Omega_x))}+|h_{\text{soil}}-h_{\text{bot}}|\|\partial_t\bar{p}\|_{L^2(0,T,(H^1(\Omega))')}\\
&+\kappa_+\tilde{K}^+\Big(\frac{\|\nabla\bar{p}\|_{L^2(\Omega_T)}}{\rho\,g}+mes(\Omega)^{1/2}\Big)+\|F_R\|_{L^2((0,T)\times\Omega_x)}\Big)\\
&\leq\frac{2}{S_0}\Big(\frac{\tilde{K}^+}{\delta}C_u+\ell\,C_p'+\kappa_+\tilde{K}^+\Big(\frac{C_p}{\rho\,g}+mes(\Omega)^{1/2}\Big)+\|F_R\|_{L^2((0,T)\times\Omega_x)}\Big) := C_u'.
\end{aligned}$$

Step 2. Nonnegativity of the solutions.

Let us solely prove that $0\leq u(t,x)$ for all $t\in(0,T)$ and for almost every $x\in\Omega_x$. Let $u_m=\sup(0,-u)$. The function u_m belongs to $L^2(0,T;V(\Omega_x))$, and is such that $\nabla u_m=-\chi_{\{u<0\}}\nabla u$ (see [6] Lemma 2.1; χ_A denotes the characteristic function of a set A). Let $\tau\in(0,T)$, setting $w(t,x)=-u_m(x,t)\chi_{(0,\tau)}(t)$ in (4.11) results in

$$S_0\int_0^\tau\langle\partial_t u,-u_m\rangle_{V',V}+\frac{\tilde{K}^-}{\ell}\int_0^\tau\int_\Omega\chi_{\{u<0\}}|\nabla u|^2 \leq \int_0^\tau\Big(\langle\int_{\bar{h}(t,x)}^{h_{\text{soil}}}\partial_t\bar{p}\,dz,-\chi(u)\,u_m\rangle_{H^1(\Omega_x)',H^1(\Omega_x)} + \int_{\Omega_x}\chi(u)\Big(F_R\,u_m-\int_{\bar{h}(t,x)}^{h_{\text{soil}}}\bar{q}dz\cdot\nabla' u_m\Big)dx\Big)dt. \tag{4.13}$$

In order to evaluate the first term in the left hand side of (4.13), we apply Lemma 2 with function f defined by $f(\lambda)=\max(0,-\lambda)$, $\lambda\in\mathbb{R}$. Of course $u_m(t,x)\neq 0$ iff $u(t,x)<0$. We have

$$\int_0^\tau\langle\partial_t u,-u_m\rangle_{V',V}dt=\frac{1}{2}\int_\Omega\big(u_m^2(\tau,x)-u_m^2(0,x)\big)\,dx=\frac{1}{2}\int_\Omega u_m^2(\tau,x)\,dx.$$

Since $\chi(u)\chi_{\{u<0\}} = 0$ by definition of χ, all the terms in the right hand side of (4.13) are null. Hence, (4.13) gives

$$\frac{S_0}{2}\int_\Omega u_m^2(\tau,x)\,dx \le -\int_0^\tau\int_\Omega \frac{\tilde{K}^-}{\ell}\chi_{\{u<0\}}|\nabla u|^2\,dx\,dt \le 0$$

and $u_m = 0$ a.e. in Ω_T.

Step 3. Uniqueness

The uniqueness of the solution is obvious since the solution u is nonnegative. Indeed, if u_1 and u_2 are two solutions of (4.9)-(4.10), then $u = u_1 - u_2$ satisfies

$$S_0\partial_t u - \nabla'\cdot\Big(\frac{\mathscr{T}^{-1}(\bar{u})}{\mathscr{T}'(\mathscr{T}^{-1}(\bar{u}))}\tilde{J}\nabla' u\Big) = 0 \quad \text{in} \quad (0,T)\times\Omega_x,$$
$$\nabla u\cdot\vec{\nu} = 0 \quad \text{on} \quad (0,T)\times\partial\Omega_x, \qquad u(0,x) = 0 \quad \text{in} \quad \Omega_x.$$

Following the previous computations, we infer from Gronwall lemma that $u = 0$ a.e. in $(0,T)\times\Omega_x$. This ends the proof of Lemma 4.4. □

The results stated in the lemma 3.1 require having regular noncylindrical domains in particular with sufficiently regular boundaries (of class $\mathscr{C}^1$ by pieces as mentioned by Mignot). Since in our problem, we cannot guarantee as much regularity at the interface h (which is in $W(0,T,\Omega_x)$), we use a regularization process to place our study within the framework of Mignot [16].

We thus regularize h by convolution in space. Let $\psi \in C^\infty(\mathbb{R}^2)$, $\psi \ge 0$, with support in the unit ball such that $\int_{\mathbb{R}^2}\psi(x)dx = 1$. For $\eta > 0$ small enough, we set $\psi_\eta(x) = \psi(x/\eta)/\eta^2$. We extend h by zero outside Ω_x, so we have $h \in C([0,T];L^2(\mathbb{R}^2))\cap W(0,T,\mathbb{R}^2)$. Hence we define $\tilde{h}$ by the convolution product with respect to the space variable

$$\tilde{h} = \psi_\eta * h.$$

Its restriction to Ω_x is denoted in the same way. It fulfills $\tilde{h} \in C^\infty(\bar{\Omega}_x)$, and as $\eta \to 0$, we have

$$\tilde{h} \to h \text{ strongly in } C([0,T];L^2(\Omega_x))\cap L^2(0,T,H^1(\Omega_x)).$$

In Eqs. (4.1)-(4.2), we replace h by $\tilde{h}$ (the substitution appears in the space integration domain Ω_t).

Let $\bar{p} \in W_1$ and $\tilde{h}\,(=\psi_\eta * h) \in C^\infty(\bar{\Omega}_x)$ where h is given by Lemma 4.4.

We thus consider the following linearized and regularized problem in Ω_T : Find $p_\eta \in W(0,T,\Omega)$ s.t. $\forall\phi \in \mathscr{A}(\mathscr{O})$ (null on the interface Γ_t defined by $\tilde{h}$)

$$\langle\partial_t p_\eta,\phi\rangle_{F',F} + \int_0^T\Big(\int_{\Omega_t}\big(\tau(\bar{p})\,B\nabla p_\eta + \kappa(\mathscr{P}^{-1}(\bar{p}))B\,\vec{e}_3\big)\cdot\nabla\phi\Big)\,dt = 0,$$
$$= \int_0^T\int_{\Omega_x}(F - aP_{|\Gamma_{\text{soil}}})\,\phi\,dx\,dt, \tag{4.14}$$
$$p_\eta = 0 \text{ in } \Omega\setminus\Omega_t,\ \forall t\in[0,T] \text{ and } p_\eta(0,x,z) = \mathscr{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0. \tag{4.15}$$

Proposition 4.5. *For any $\eta > 0$, there exists a unique function p_η in $W(0,T,\Omega)$ solution of (4.14)-(4.15). It fulfills the uniform estimates*

$$\|p_\eta\|_{L^2(0,T;H^1(\Omega))} \le C_p \quad \text{and} \quad \|p_\eta\|_{L^2(0,T;(H^1(\Omega))')} \le C_p', \tag{4.16}$$

where C_p and C_p' only depend on the data of the original problem (3.17)- (3.18).

Let us admit for the moment this Proposition whose the proof will be given at the end. From now, we omit the subscript η in p_η (and then in u_η).

Let $(\bar{u},\bar{p}) \in W(0,T,\Omega)\times W(0,T,\Omega_x)$ the unique solution of (4.9)-(4.10) and (4.14)-(4.15), Lemma 4.4 and Proposition 4.5 enable to define an application $\mathscr{F}$ such that:

$$W(0,T,\Omega_x)\times W(0,T,\Omega) \to W(0,T,\Omega_x)\times W(0,T,\Omega)$$
$$\mathscr{F}(\bar{u},\bar{p}) = (u,p). \tag{4.17}$$

The end of the present subsection is devoted to the proof of the existence of a fixed point of $\mathscr{F}$ in some appropriate subset. We conclude the proof of Theorem 3.7 by passing to the limit when $\eta \to 0$.

Lemma 4.6. *Let $(\bar{u},\bar{p}) \in W(0,T,\Omega)\times W(0,T,\Omega_x)$ the unique solution of (4.9)-(4.10) and (4.14)-(4.15), thus*

- *There exists $\mathscr{C}$ a nonempty, closed, convex, bounded set in $W(0,T,\Omega_x)\times W(0,T,\Omega)$ satisfying $\mathscr{F}(\mathscr{C})\subset\mathscr{C}$,*

- *The application $\mathscr{F}$ defined by (4.17) is weakly sequentially continuous in $W(0,T,\Omega_x)\times W(0,T,\Omega)$,*
- *There exists $(u,p)\in W_1\times W_2$ such that $\mathscr{F}(u,p)=(u,p)$.*

Proof. We set $\mathscr{C}=W_1\times W_2$ where W_1 and W_2 are defined in (4.7)- (4.8). The first point of Lemma 4.6 is obvious thanks to Lemma 4.4 and Proposition 4.5. Indeed $\mathscr{C}$ is clearly a nonempty (strongly) closed convex set in $W(0,T,\Omega_x)\times W(0,T,\Omega)$.
Regarding the second point of Lemma 4.6, we first note that $\mathscr{C}$ is compact for the weak topology. $\mathscr{F}$ maps $W_1\times W_2$ into it self. Let now $(v_n)_{n\geq 0}=(\bar{u}_n,\bar{p}_n)_{n\geq 0}$ be any sequence in $\mathscr{C}$ which is weakly convergent in $W(0,T,\Omega_x)\times W(0,T,\Omega)$, and let $v=(\bar{u},\bar{p})$ be its weak limit. We aim to show that

$$\mathscr{F}(v_n)\rightharpoonup\mathscr{F}(v)\quad \text{in } W(0,T,\Omega_x)\times W(0,T,\Omega) \text{ as } n\to\infty.$$

Since $\mathscr{F}(v_n)\in W_1\times W_2$ and $W_1\times W_2$ is weakly compact, it is sufficient to show that there exits a subsequence (v_n') of (v_n) such that $\mathscr{F}(v_n')\rightharpoonup\mathscr{F}(v)$. Extracting a subsequence if needed we may assume without loss of generality that $\mathscr{F}(v_n)\rightharpoonup w$ in $W(0,T,\Omega_x)\times W(0,T,\Omega)$ as $n\to\infty$ for some $w=(u,p)\in W_1\times W_2$, and we have to show that w and $\mathscr{F}(v)$ agree. Set $w_n=\mathscr{F}(v_n)$ $(w_n=(u_n,p_n))$, it follows from Aubin's Lemma that

$$\begin{aligned}
w_n\to w \quad &\text{in} \quad L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega) \quad \text{and} \quad w_n(t,x)\to w(t,x) \quad \text{a.e.;}\\
v_n\to v \quad &\text{in} \quad L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega) \quad \text{and} \quad v_n(t,x)\to v(t,x) \quad \text{a.e.;}\\
&\partial_t w_n\rightharpoonup\partial_t w \quad \text{in} \quad L^2(0,T;H^1(\Omega_x)')\times L^2(0,T;H^1(\Omega)')\\
&\nabla w_n\rightharpoonup\nabla w \quad \text{weakly in} \quad L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega).
\end{aligned}$$

Thanks to Lebesgue theorem (and the properties of functions κ,τ and T_l) we obtain that $w=\mathscr{F}(v)$ (since $w(0,.)=(u(0,.),p(0,.))=(u_0,p_0)$ because $w\in\mathscr{C}$) and the proof that $\mathscr{F}|_{\mathscr{C}}$ be weakly sequentially continuous is complete.
It follows from Schauder theorem [22] that there exists $(u,p)\in W_1\times W_2$ such that $\mathscr{F}(u,p)=(u,p)$. The proof of Lemma 4.6 is thus achieved. □

We collect the results obtained previously. We can associate with any real number $\eta>0$ the fixed point point $(u_\eta,p_\eta)\in W_1\times W_2$ of the mapping $\mathscr{F}$. It is a solution of the system :

$$\partial_t p_\eta-\nabla\cdot\big(\tau(p_\eta)\,B\nabla p_\eta\big)-\nabla\cdot\Big(\kappa(\mathscr{P}^{-1}(p_\eta))B\,\vec{e}_3\Big)=0\quad\text{in}\quad\mathscr{O}_T, \tag{4.18}$$

$$p_{\eta|_{\Gamma_t}}=\mathscr{P}(P_s)\quad\text{in}\quad(0,T),\qquad\nabla\big(\mathscr{P}^{-1}(p_\eta)+\rho g z\big)\cdot\vec{\nu}=0\quad\text{on}\quad(0,T)\times\Gamma_{ver},$$

$$a\,P+\nabla\big(\mathscr{P}^{-1}(p_\eta)+\rho g z\big)\cdot\vec{\nu}=F\quad\text{on}\quad(0,T)\times\Gamma_{\text{soil}},\qquad p_\eta(0,x,z)=\mathscr{P}(P_0)(x,z)\,\text{in}\,\Omega_0. \tag{4.19}$$

$$S_0\partial_t u_\eta-\nabla'\cdot\big(\tilde{J}(\mathscr{T}^{-1}(u_\eta))\nabla'\mathscr{T}^{-1}(u_\eta)\big)=-\chi(u_\eta)\Big(\int_{h_\eta(t,x)}^{h_{\text{soil}}}\big(\partial_t p_\eta\big)\,dz \tag{4.20}$$

$$+\operatorname{div}_x\Big(\int_{h_\eta(t,x)}^{h_{\text{soil}}}q_\eta\,dz\Big)-F_R\Big)\,\text{in}\,(0,T)\times\Omega_x, \tag{4.21}$$

$$\nabla u_\eta\cdot\vec{\nu}=0\quad\text{on}\quad(0,T)\times\partial\Omega_x,\quad u_\eta(0,x)=\frac{(h_0(x)-h_{\text{bot}}(x))^2}{2}\quad\text{in}\quad\Omega_x. \tag{4.22}$$

We can obtain similar estimates for (u_η,p_η) than those derived in Lemma 4.4 and Proposition 4.5. We thus assert the existence of limit functions (extracting a subsequence if needed) $(u,p)\in W(0,T,\Omega_x)\times W(0,T,\Omega)$ such that

$$\begin{aligned}
(u_\eta,p_\eta)&\to(u,p) && \text{in} && L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega)\\
(u_\eta(t,x),p_\eta(t,x))&\to(u(t,x),p(t,x)) && \text{a.e} && \text{in} \quad ((0,T)\times\Omega_x)\times((0,T)\times\Omega)\\
\tilde{h}(t,x)=\psi_\eta*h(t,x)&\to h(t,x), && \text{a.e} && \text{in} \quad (0,T)\times\Omega_x\\
(\partial_t u_\eta,\partial_t p_\eta)&\rightharpoonup(\partial_t u,\partial_t p) && \text{in} && L^2(0,T;H^1(\Omega_x)')\times L^2(0,T;H^1(\Omega)')\\
(\nabla u_\eta,\nabla p_\eta)&\rightharpoonup(\nabla u,\nabla p) && \text{weakly in} && L^2((0,T)\times\Omega_x)\times L^2((0,T)\times\Omega).
\end{aligned}$$

Letting $\eta\to 0$ in weak formulations resulting from (4.18)-(4.22), we prove the existence of a weak solution (u,p) of problem (3.17)-(3.20). This ends the proof of Theorem 3.7. □

4.2. Proof of Proposition 4.5

Again, we omit the subscript η in p_η.

The proof of Proposition 4.5 is done in two steps. We first use the method of auxiliary domains presented in [16] to solve difficulties related to the free boundary. That is, we extend the functions out of the domain of study by zero, then we introduce a penalized problem and go to the limit to return to the linearized problem (4.14)-(4.15).

We thus consider the weak solution p of the linearized problem (4.14)-(4.15).

So, $\forall \phi \in L^2([0,T], H^1(\Omega_t))(= H^{0,1}(\mathscr{O}_T))$ with $\phi|_{\Gamma_t} = 0$, we look for $p \in W(0,T,\Omega) = 0$ such that

$$\langle \partial_t p, \phi\rangle_{F',F} + \int_0^T \Big(\int_{\Omega_t} \big(\tau(\bar{p}) B \nabla p + \kappa(\mathscr{P}^{-1}(\bar{p})) B \vec{e}_3\big) \cdot \nabla\phi \big) d\boldsymbol{x}\Big) dt = \int_0^T \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}})\, \phi_{|\Gamma_{\text{soil}}}\, dx\, dt.$$

We first remark that the solution of system (4.14)-(4.15) is unique. Indeed, if p_1 and p_2 are two solutions of (4.14)-(4.15), then $d = p_1 - p_2$ satisfies

$$\langle \partial_t d, \phi\rangle_{F',F} + \int_0^T \Big(\int_{\Omega_t} \big(\tau(\bar{p}) B \nabla d \cdot \nabla\phi\big) d\boldsymbol{x}\Big) dt = 0.$$

Then, taking $\phi = d$ and using the fourth point of Lemma 3.1, we conclude that

$$\frac{1}{2}\int_{\Omega_T} d^2(T,x)\, d\boldsymbol{x}\, dt + \tau_- K^- \int_0^T \int_{\Omega_t} |\nabla d|^2\, d\boldsymbol{x}\, dt \le 0,$$

since $d(0,.) = 0$. We infer from this equality that $d = 0$ a.e. in $(0,T)\times\Omega$ (since $d = 0$ on the interface Γ_T).

We will define a family of approximate problems which are linear parabolic problems in the cylindrical domain $(0,T)\times\Omega$, and whose the solution restricted to the set $\mathscr{O}_T$ will converge to the p solution of the linearized equation (4.14).

Step 1. Penalized problems

Let $\epsilon > 0$, we now consider the following penalized problem on Ω : Find $p_\epsilon \in W(0,T,\Omega)$ s.t. $\forall \phi \in L^2(0,T;\mathscr{D}(\Omega))$

$$\begin{aligned}
&\langle \partial_t p_\epsilon, \phi\rangle_{F',F} + \int_0^T \Big(\int_\Omega \big(\tau(\bar{p}) \tilde{N} \nabla p_\epsilon + \kappa(\mathscr{P}^{-1}(\bar{p})) \tilde{N}\vec{e}_3\big) \cdot \nabla\phi\big)\Big)\, d\boldsymbol{x}\, dt \\
&- \int_0^T \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}})\, \phi_{|\Gamma_{\text{soil}}}\, dx\, dt + \int_{\mathscr{O}'_T} \nabla p_\epsilon \cdot \nabla\phi\, d\boldsymbol{x}\, dt + \frac{1}{\epsilon}\int_{\mathscr{O}'_T} p_\epsilon \phi\, d\boldsymbol{x}\, dt = 0, \qquad (4.23)\\
&p_\epsilon(0,x,z) = \mathscr{P}(P_0)(x,z) \quad \text{in} \quad \Omega_0 \quad \text{and} \quad p_\epsilon(0,x,z) = 0 \quad \text{in} \quad \Omega\setminus\Omega_0. \qquad (4.24)
\end{aligned}$$

where $\tilde{N} = B$ in $\mathscr{O}_T$ and $\tilde{N} = 0$ in $\mathscr{O}'_T$.

We aim to state that the penalized system (4.23)-(4.24) admits a unique solution p_ϵ which tends to the solution of problem (4.14)-(4.15) when $\epsilon \to 0$. Eq. (4.23) can be written

$$\begin{aligned}
&\langle \partial_t p_\epsilon, \phi\rangle_{F',F} + \underbrace{\int_0^T \int_\Omega \tau(\bar{p})\, \tilde{N}\nabla p_\epsilon \cdot \nabla\phi\, d\boldsymbol{x}\, dt + \int_{\mathscr{O}'_T} \nabla p_\epsilon \cdot \nabla\phi\, d\boldsymbol{x}\, dt + \frac{1}{\epsilon}\int_{\mathscr{O}'_T} p_\epsilon\phi\, d\boldsymbol{x}\, dt}_{A_\epsilon(p_\epsilon,\phi)} \\
&= \underbrace{-\int_0^T \Big(\int_\Omega \big(\kappa(\mathscr{P}^{-1}(\bar{p}))\tilde{N}\vec{e}_3 \cdot \nabla\phi\big)\Big) d\boldsymbol{x} - \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}})\, \phi_{|\Gamma_{\text{soil}}}\, dx\Big) dt}_{L_\epsilon(\phi)} \quad \forall \phi \in W(0,T,\Omega). \qquad (4.25)
\end{aligned}$$

Due to (3.10), we establish that the coefficients of A_ϵ are in $L^\infty((0,T)\times\Omega)$. Moreover, we have

$$A_\epsilon(p,p) \ge inf\big(1, \tau_- K^-, \frac{1}{\epsilon}\big)\, \|p\|_{L^2(0,T,H^1(\Omega))}, \quad \forall p \in L^2(0,T,H^1(\Omega)).$$

We directly check that L_ϵ is a linear form on $L^2(0,T,H^1(\Omega))$. We thus deduce the existence and uniqueness for the system (4.23)-(4.24).

STEP 2. LIMIT WHEN $\epsilon \to 0$

We first derive some uniforme estimates with respect to ϵ (and also η). Multiplying Eq. (4.25) by p_ϵ and integrating by parts over Ω, we thus obtain $\forall s \leq T$

$$\langle \partial_t p_\epsilon, p_\epsilon \rangle_{F',F} + \underbrace{\int_0^s \int_{\Omega_t} \tau(\bar{p}) \tilde{N} |\nabla p_\epsilon|^2 \, d\boldsymbol{x}\, dt + \int_{\mathscr{O}'_s} |\nabla p_\epsilon|^2 \, d\boldsymbol{x}\, dt + \frac{1}{\epsilon} \int_{\mathscr{O}'_s} p_\epsilon^2 \, d\boldsymbol{x}\, dt}_{I_1}$$
$$= \underbrace{-\int_0^s \Big(\int_\Omega \frac{\rho g}{\mu} \kappa(\mathscr{P}^{-1}(\bar{p})) \tilde{N} \vec{e}_3 \cdot \nabla p_\epsilon \, d\boldsymbol{x} - \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}}) \, p_{\epsilon_{|\Gamma_{\text{soil}}}} \, dx \Big) dt}_{I_2}.$$

Then, applying Lemma 3.1 to the first term, we get

$$|\langle \partial_t p_\epsilon, p_\epsilon \rangle_{F',F}| = |\int_0^s \langle \partial_t p_\epsilon, p_\epsilon \rangle_{V'(\Omega),V(\Omega)} \, dt| = \frac{1}{2} \Big(\int_\Omega p_\epsilon^2(s,.) d\boldsymbol{x} - \int_\Omega p_\epsilon^2(0,.) d\boldsymbol{x} \Big)$$

Besides

$$|I_1| \geq \tau_- K^- \|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega_t))} + \|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega'_t))} + \frac{1}{\epsilon} \int_{\mathscr{O}'_s} p_\epsilon^2 \, d\boldsymbol{x}\, dt,$$
$$|I_2| \leq \frac{\epsilon_1}{4} \int_0^s \int_\Omega p_\epsilon^2(s,.) d\boldsymbol{x} + \frac{\epsilon_1}{2} \|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega_t))}$$
$$+ \frac{1}{\epsilon_1} T \, mes(\Omega) \Big(\frac{\rho g \kappa_+ K^+}{\mu} \Big)^2 + \frac{1}{\epsilon_1} \| \underbrace{F - aP_{|\Gamma_{\text{soil}}}}_{F_R} \|^2_{L^2([0,T],L^2(\Omega_x))}.$$

By taking $\epsilon_1 = \tau_- K^-$, thanks to Gronwall's Lemma, we deduce that there exists a constant C_p depending only on the data such that

$$\|p_\epsilon\|^2_{L^2([0,T],H^1(\Omega_t))} \quad \leq \quad C_p. \tag{4.26}$$

More precisely, we have

$$\tau_- K^- \|\nabla p_\epsilon\|^2_{L^2([0,T],L^2(\Omega_t))} \quad \leq \quad C_0 \big(1 + \frac{\tau_- K^- T}{2} e^{\frac{\tau_- K^- T}{2}}\big),$$

where

$$C_0 = \int_\Omega p_0^2 d\boldsymbol{x} + \frac{2}{\tau_- K^-} \big(T \, mes(\Omega) \Big(\frac{\rho g \kappa_+ K^+}{\mu} \Big)^2 + \|F_R\|^2_{L^2([0,T],L^2(\Omega_x))} \big).$$

So the sequence $\{p_\epsilon\}$ is bounded in $L^2(0,T;H^1(\Omega))$ and the sequence $\{\frac{1}{\sqrt{\epsilon}} p_\epsilon\}$ is bounded in $L^2((0,T) \times \Omega'_T)$. We can thus extract subsequences $\{p_\epsilon\}$, $\{\frac{1}{\sqrt{\epsilon}} p_\epsilon\}$ (not relabeled for convenience) and there exist $r \in L^2(0,T;H^1(\Omega))$ and $r' \in L^2((0,T) \times \Omega'_T)$ such that

$$p_\epsilon \rightharpoonup r \quad \text{weakly in} \quad L^2((0,T) \times \Omega) \tag{4.27}$$
$$\frac{1}{\sqrt{\epsilon}} p_\epsilon \rightharpoonup r' \quad \text{weakly in} \quad L^2((0,T) \times \Omega'_T) \tag{4.28}$$
$$\nabla p_\epsilon \rightharpoonup \nabla p \quad \text{weakly in} \quad L^2((0,T) \times \Omega). \tag{4.29}$$

It results from the first two convergences that

$$p_\epsilon|_{\mathscr{O}'_T} \rightharpoonup r|_{\mathscr{O}'_T} \quad \text{weakly in} \quad L^2((0,T) \times \Omega'_T)$$
$$p_\epsilon|_{\mathscr{O}'_T} = \sqrt{\epsilon} \times \frac{1}{\sqrt{\epsilon}} p_\epsilon \rightharpoonup 0 \quad \text{weakly in} \quad L^2((0,T) \times \Omega'_T),$$

so

$$r|_{\mathscr{O}'_T} = 0. \tag{4.30}$$

Moreover, since $r \in F([0,T] \times \Omega)$, we infer from the second result of Lemma 3.1 that we can define $\gamma(r)$ on Γ_t and $\gamma(r) = 0$ on Γ_t for all $0 \leq t \leq T$ thanks to (4.30). Thus $r|_{\mathscr{O}_T} \in F(\mathscr{O}_T)$.

We must now check that $r|_{\mathscr{O}_T}$ satisfies (4.14). Let $\phi \in F(\mathscr{O}_T)$ that we extend by zero on $\mathscr{O}'_T$, the extension is still denoted by ϕ. Thus taking $\phi \in F((0,T)\times\Omega)$ in (4.25), we get

$$\int_0^T \langle \partial_t p_\epsilon, \phi\rangle_{V',V}\, dt + \int_0^T \left(\int_\Omega \left(\tau(\bar{p})\, B\nabla p_\epsilon + \kappa(\mathscr{P}^{-1}(\bar{p}))B\vec{e}_3\right)\cdot\nabla\phi\right) d\boldsymbol{x}\, dt = \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}})\,\phi_{|\Gamma_{\text{soil}}}\, dx,$$

that we can write as follows by choosing $\phi \in B_T(\mathscr{O}_T)$

$$-\int_0^T \langle p_\epsilon, \partial_t\phi\rangle_{V',V}\, dt + \int_0^T \left(\int_\Omega \left(\tau(\bar{p})\, B\nabla p_\epsilon + \kappa(\mathscr{P}^{-1}(\bar{p}))B\vec{e}_3\right)\cdot\nabla\phi\right) d\boldsymbol{x}\, dt = \int_{\Omega_0} p_0(\boldsymbol{x})\phi_0(0,\boldsymbol{x})\, d\boldsymbol{x} + \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}})\,\phi_{|\Gamma_{\text{soil}}}\, dx.$$

By letting $\epsilon \to 0$, we obtain

$$-\int_0^T \langle r|\mathscr{O}_T, \partial_t\phi\rangle_{V',V}\, dt + \int_0^T \left(\int_{\Omega_t} \left(\tau(\bar{p})B\nabla r_{|\mathscr{O}_T} + \kappa(\mathscr{P}^{-1}(\bar{p}))B\vec{e}_3\right)\cdot\nabla\phi\right) d\boldsymbol{x}\, dt = \int_{\Omega_0} p_0(\boldsymbol{x})\phi_0(0,\boldsymbol{x})\, d\boldsymbol{x} + \int_{\Omega_x} (F - aP_{|\Gamma_{\text{soil}}})\,\phi_{|\Gamma_{\text{soil}}}\, dx. \quad (4.31)$$

It remains to be established that

$$D_t(r|\mathscr{O}_T) \in F'(\mathscr{O}_T) \quad \text{and} \quad D_t(p_\epsilon|\mathscr{O}_T) \rightharpoonup D_t(r|\mathscr{O}_T) \quad \text{in} \quad F'(\mathscr{O}_T).$$

From (4.31), we deduce that, taking $\phi \in \mathscr{D}(\mathscr{O}_T)$

$$D_t(r|\mathscr{O}_T) - \nabla\cdot\left(\tau(\bar{p})\, B\nabla q|\mathscr{O}_T + \kappa(\mathscr{P}^{-1}(\bar{p}))B\vec{e}_3\right) = 0 \quad \text{in} \quad \mathscr{D}'(\mathscr{O}_T).$$

Since $D'(\mathscr{O}_T)$ is dense in $F'(\mathscr{O}_T)$, the previous equality holds true in $F'(\mathscr{O}_T)$. Moreover the solution $p_\epsilon|\mathscr{O}_T$ of (4.25) verifies

$$D_t(p_\epsilon|\mathscr{O}_T) - \nabla\cdot\left(\tau(\bar{p})\, B\nabla p_\epsilon|\mathscr{O}_T + \kappa(\mathscr{P}^{-1}(\bar{p}))B\vec{e}_3\right) = 0 \quad \text{in} \quad F'(\mathscr{O}_T).$$

Letting $\epsilon \to 0$ and we infer from convergences (4.27)-(4.29) that

$$\underbrace{\nabla\cdot\left(\tau(\bar{p})\, B\nabla p_\epsilon|_{\mathscr{O}_T} + \kappa(\mathscr{P}^{-1}(\bar{p}))B\vec{e}_3\right)}_{=D_t(p_\epsilon|_{\mathscr{O}_T})} \rightharpoonup \underbrace{\nabla\cdot\left(\tau(\bar{p})\, B\nabla r|_{\mathscr{O}_T} + \kappa(\mathscr{P}^{-1}(\bar{p}))B\vec{e}_3\right)}_{=D_t(r|_{\mathscr{O}_T})},$$

weakly in $F'(\mathscr{O}_T)$. Thus $r|\mathscr{O}_T$ is the unique solution of (4.14)-(4.15), and the limit of $p_\epsilon|_{\mathscr{O}_T}$ being independent of the chosen subsequence, the whole sequence converges towards $r|\mathscr{O}_T$ in $\mathscr{A}(\mathscr{O}_T)$. Moreover, we obtain the first part of (4.16) for the solution $r \in L^2(0,T,H^1(\Omega))$ of system (4.14)-(4.15) in the same way that for estimate (4.26) obtained for p_ϵ. Finally, as was done in Lemma 4.4, we deduce from the first inequality of (4.16) that

$$\|\partial_t r\|^2_{L^2(0,T,H^1(\Omega)')} \leq C'_p,$$

where C'_p depends on the data and on C_p. This ends the proof of Proposition 4.5. □

References

[1] Philippe Ackerer and Anis Younes. Efficient approximations for the simulation of density driven flow in porous media. *Advances in Water Resources*, 31(1):15 – 27, 2008.
[2] H. W. Alt and E. DiBenedetto. Nonsteady flow of water and oil through inhomogeneous porous media. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 12(3):335–392, 1985.
[3] Hans Wilhelm Alt and Stephan Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.*, 183(3):311–341, 1983.
[4] Jacob Bear. *Dynamics of Fluids in Porous Media*. Elsevier, New-York, 1972.
[5] Jacob Bear and Arnold Verruijt. *Modeling Groundwater Flow and Pollution*. Springer, Netherlands, 1987.
[6] Philippe Bénilan, Lucio Boccardo, Thierry Gallouët, Ron Gariepy, Michel Pierre, and Juan Luis Vázquez. An L^1-theory of existence and uniqueness of solutions of nonlinear elliptic equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 22(2):241–273, 1995.
[7] Christophe Bourel, Catherine Choquet, Carole Rosier, and Munkhgerel Tsegmid. Modeling of shallow aquifers in interaction with overland water. *Appl. Math. Model.*, 81:727–751, 2020.
[8] Xinfu Chen, Avner Friedman, and Tsuyoshi Kimura. Nonstationary filtration in partially saturated porous media. *European J. Appl. Math.*, 5(3):405–429, 1994.
[9] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.
[10] Gérard Gagneux and Monique Madaune-Tort. *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière*, volume 22 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1996. With a preface by Charles-Michel Marle.

[11] Josephus Hulshof and Noemí Wolanski. Monotone flows in N-dimensional partially saturated porous media: Lipschitz-continuity of the interface. *Arch. Rational Mech. Anal.*, 102(4):287–305, 1988.
[12] M. Jazar and R. Monneau. Derivation of seawater intrusion models by formal asymptotics. *SIAM J. Appl. Math.*, 74(4):1152–1173, 2014.
[13] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva. *Linear and quasilinear equations of parabolic type*. Translated from the Russian by S. Smith. Translations of Mathematical Monographs, Vol. 23. American Mathematical Society, Providence, R.I., 1968.
[14] J.-L. Lions. Sur les problèmes mixtes pour certains systèmes paraboliques dans des ouverts non cylindriques. *Ann. Inst. Fourier (Grenoble)*, 7:143–182, 1957.
[15] J.-L. Lions and E. Magenes. *Problèmes aux limites non homogènes et applications. Vol. 1*. Travaux et Recherches Mathématiques, No. 17. Dunod, Paris, 1968.
[16] Alain Lucien Mignot. Méthodes d'approximation des solutions de certains problèmes aux limites linéaires. I. *Rend. Sem. Mat. Univ. Padova*, 40:1–138, 1968.
[17] Safa Al Nazer, Carole Rosier, and Munkhgerel Tesgmid. Derivation and mathematical analysis of dupuit-richards model taking into account the fluid compressibility. *submitted*, 2020.
[18] Gary Pantelis. Saturated-unsaturated flow in unconfined aquifers. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 36(5):648–657, Sep 1985.
[19] R. E. Showalter and Ning Su. Partially saturated flow in a poroelastic medium. *Discrete Contin. Dyn. Syst. Ser. B*, 1(4):403–420, 2001.
[20] Jacques Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl. (4)*, 146:65–96, 1987.
[21] Hong-Ming Yin. A singular-degenerate free boundary problem arising from the moisture evaporation in a partially saturated porous medium. *Ann. Mat. Pura Appl. (4)*, 161:379–397, 1992.
[22] Eberhard Zeidler. *Nonlinear functional analysis and its applications. I*. Springer-Verlag, New York, 1986. Fixed-point theorems, Translated from the German by Peter R. Wadsack.

Appendix

We start by recalling a general way to take into account on the small compressibility of the fluid in the 3d-Richards problem. This also can be found in [17].

A.1. 3d-Richards equations for compressible fluid

Conservation laws. The basis of the modeling is the mass conservation law coupled with the classical Darcy law for porous media. Fluid is now considered to be weakly compressible.

We begin with the conservation of momentum. In view of the (large) dimensions of an aquifer (related to the characteristic size of the porous structure of the underground), we consider a continuous description of the porous medium. We denote by ρ_0 the reference density of the fluid. The effective velocity q of the flow is thus related to the pressure P through the Darcy law associated with a non-linear anisotropic conductivity

$$q = -\frac{\kappa K_0}{\mu}(\nabla P + \rho g \nabla z),$$

where ρ and μ are respectively the density and the viscosity of the fluid, K_0 is the permeability of the soil, $\kappa(P)$ is the relative conductivity and g the gravitational acceleration constant. Introducing the hydraulic head H defined by

$$H = \frac{P}{\rho_0 g} + z, \tag{A.32}$$

we write the previous equation as follows:

$$q = -\kappa K \nabla H - \frac{\kappa(P) K_0}{\mu}(\rho - \rho_0) g \nabla z, \quad K = \frac{K_0 \rho_0 g}{\mu}. \tag{A.33}$$

In this relation, the matrix κK is the hydraulic conductivity which expresses the ability of the underground to conduct the fluid.

Next, the conservation of mass during displacement is given by the following equation

$$\partial_t(\theta\rho) + \nabla \cdot (\rho q) = 0. \tag{A.34}$$

The function θ is the volumetric moisture content defined by

$$\theta = \phi s,$$

where ϕ is the porosity of the medium and s is the saturation. If we assume that the air present in the unsaturated zone has infinite mobility, the saturation s and then the function θ are thus considered as monotone functions depending on the pressure as we will detail latter.

State equation for the fluid compressibility. We consider that the fluid are compressible by assuming that pressure P is related to the density ρ such that $\frac{d\rho}{dP} = \rho\alpha_P$, that is:

$$\rho = \rho_0 e^{\alpha_P(P-P_0)}. \tag{A.35}$$

The real number $\alpha_P \geq 0$ is the fluid compressibility coefficient and P_0 is the pressure of reference . Further assuming $\alpha_P = 0$ we would recover the incompressible case.

Compressibility of the fluid. First, we assume that the fluid (namely here fresh water) is weakly compressible. It means that

$$\alpha_P \ll 1. \tag{A.36}$$

Let us exploit this assumption. In natural conditions and especially in an aquifer, one observes small fluid mobility (defined by the ratio κ/μ). First consequence of the low compressibility of the fluid combined with the low mobility of fluid appears in the momentum equation. We perform a Taylor expansion with regard to P of the density ρ in the gravity term of the Darcy equation. Neglecting the terms weighted by $\alpha_P\kappa/\mu \ll 1$ in (A.33), we get:

$$q = -\kappa K\nabla H, \quad K = \frac{\rho_0 g}{\mu}K_0. \tag{A.37}$$

Second consequence is $\nabla\rho\cdot q \ll 1$ which leads to the following simplification in the mass conservation equation (A.34):

$$\rho\partial_t\theta + \theta\,\partial_t\rho + \rho\nabla\cdot q = 0.$$

Neglecting in this way the variation of density in the direction of flow is sometimes considered as an extra assumption called Bear's hypothesis (*cf* [1]). Including (A.35), that is $\partial_t\rho = \rho\alpha_P\partial_t P$ in the latter equation, we get

$$\rho\partial_t\theta + \rho\,\theta\alpha_P\,\partial_t P + \rho\nabla\cdot q = 0.$$

After simplification by $\rho > 0$, we finally obtain

$$\partial_t\theta + \theta\alpha_P\,\partial_t P + \nabla\cdot q = 0. \tag{A.38}$$

Equivalently, using the hydraulic head (A.32) and the Darcy law (A.37), (A.38) can be written

$$\partial_t\theta + S_0\partial_t H - \nabla\cdot(K\nabla H) = 0 \quad \text{where} \quad S_0 = \rho_0\, g\,\phi\,\alpha_P. \tag{A.39}$$

We notice that if the fluid is assumed incompressible, $\alpha_P = 0$, then Eq. (A.38) is the classical Richards equation in pressure formulation. An adequate definition of the volumetric moisture content θ and of the mobility function κ is the key of the model.

Compressible 3d-Richards problem. Finally the compressible 3d-Richards equations are obtained collecting (A.37), (A.38) and the boundary conditions:

$$\begin{cases} \partial_t\theta(P) + \theta\alpha_P\,\partial_t P + \nabla\cdot q = 0 & \text{in } (0,T)\times\Omega, \\ q = -K\nabla\left(\dfrac{P}{\rho_o\, g} + z\right) & \text{in } (0,T)\times\Omega, \\ P(0,x,z) = P_{\text{init}}(x,z) & \text{for } (x,z)\in\Omega_0. \\ + \text{boundary conditions(1.8)and(1.9)}. \end{cases} \tag{A.40}$$

In the rest of the paper, we will write ρ instead of ρ_0 when no confusion is possible.

A.2. Justification of the coupled with non-vanishing horizontal conductivity for weakly compressible fluid

This subsection, is devoted to the justification of the model (2.2)–(2.8). More precisely and following the strategy of [7], we determine the effective models associated to (2.2)–(2.8) and to the 3d-Richards problem (1.10) when the ratio *characteristic deepness/characteristic length of the shallow aquifer* tends to zero. We do that for different time scale and show in Proposition A.7 that those effectives models coincide under some hypothesis.

Dimensionless form of the 3D-Richards and coupled problem

We introduce a fixed dimensionless reference domain $\overline{\Omega}$ of type (1.2) and a dimensionless real number $\overline{T} > 0$. We fix $\overline{\Omega}_x$, $\overline{h}_{\rm soil}$, and $\overline{h}_{\rm bot}$ such that

$$\overline{\Omega} = \big\{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega}_x \times \mathbb{R} \quad | \quad \overline{z} \in \big]\overline{h}_{\rm bot}(\overline{\boldsymbol{x}}), \overline{h}_{\rm soil}(\overline{\boldsymbol{x}})\big[\big\}.$$

The boundary of $\overline{\Omega}$ is decomposed into $\overline{\Gamma}_{\rm bot} := \{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega} \mid \overline{z} = \overline{h}_{\rm bot}(\overline{\boldsymbol{x}})\}$, $\overline{\Gamma}_{\rm soil} := \{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega} \mid \overline{z} = \overline{h}_{\rm soil}(\overline{\boldsymbol{x}})\}$, and $\overline{\Gamma}_{\rm ver} := \{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega} \mid \overline{\boldsymbol{x}} \in \partial\overline{\Omega}_x\}$. To obtain a rescaled version of problems (1.10) and (2.2)–(2.8) in the domain $]0, \overline{T}[\times\overline{\Omega}$, we introduce positive reference numbers L_x, L_z, and T such that the physical variables are given as a function of the dimensionless variables by

$$\boldsymbol{x} = L_x \overline{\boldsymbol{x}}, \quad z = L_z \overline{z}, \quad t = T\overline{t}/\overline{T}.$$

The physical domain Ω may then be viewed as a dilation of the reference domain $\overline{\Omega}$. We set

$$\Omega_x = L_x \overline{\Omega}_x, \quad h_{\rm soil}(\boldsymbol{x}) = L_z\, \overline{h}_{\rm soil}(\overline{\boldsymbol{x}}), \quad h_{\rm bot}(\boldsymbol{x}) = L_z\, \overline{h}_{\rm bot}(\overline{\boldsymbol{x}}).$$

The reference exterior normal is given by

$$\overline{\boldsymbol{n}}(\overline{\boldsymbol{x}}, \overline{z}) = \begin{cases} \Big(\boldsymbol{e}_3 - (L_z/L_x)\nabla_{\overline{\boldsymbol{x}}}\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})\Big)\Big((L_z^2/L_x^2)|\nabla_{\overline{\boldsymbol{x}}}\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})|^2 + 1\Big)^{-1/2} & \text{on } \overline{\Gamma}_{\rm soil} \\ \Big((L_z/L_x)\nabla_{\overline{\boldsymbol{x}}}\overline{h}_{\rm bot}(\overline{\boldsymbol{x}}) - \boldsymbol{e}_3\Big)\Big((L_z^2/L_x^2)|\nabla_{\overline{\boldsymbol{x}}}\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})|^2 + 1\Big)^{-1/2} & \text{on } \overline{\Gamma}_{\rm bot} \\ \boldsymbol{n}(\boldsymbol{x}, z) & \text{on } \overline{\Gamma}_{\rm ver} \end{cases}$$

where the vector $\boldsymbol{n}$ is horizontal and does not change during the rescaling. We also introduce the reference subdomains $\Omega_{\overline{h}}^-$ and $\Omega_{\overline{h}}^+$ by

$$\Omega_{\overline{h}}^-(\overline{t}) = \big\{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega}_x \times \mathbb{R} \mid \overline{z} \in \big]\overline{h}_{\rm bot}(\overline{\boldsymbol{x}}), \overline{h}(\overline{t}, \overline{\boldsymbol{x}})\big[\big\}, \qquad \Omega_{\overline{h}}^+(\overline{t}) = \big\{(\overline{\boldsymbol{x}}, \overline{z}) \in \overline{\Omega}_x \times \mathbb{R} \mid \overline{z} \in \big]\overline{h}(\overline{t}, \overline{\boldsymbol{x}}), \overline{h}_{\rm soil}(\overline{\boldsymbol{x}})\big[\big\}$$

where $\overline{h}$ is the function defined by

$$L_z\overline{h}(\overline{t}, \overline{\boldsymbol{x}}) = h(t, \boldsymbol{x}).$$

The rescaled unknowns are such that

$$L_z\overline{P}(\overline{t}, \overline{\boldsymbol{x}}, \overline{z}) = P(t, \boldsymbol{x}, z), \quad \overline{\boldsymbol{v}}(\overline{t}, \overline{\boldsymbol{x}}, \overline{z}) = v(t, \boldsymbol{x}, z), \quad \overline{\boldsymbol{q}}(\overline{t}, \overline{\boldsymbol{x}}, \overline{z}) = u(t, \boldsymbol{x}, z), \quad \overline{\boldsymbol{w}}(\overline{t}, \overline{\boldsymbol{x}}, \overline{z}) = w(t, \boldsymbol{x}, z),$$
$$L_z\overline{H}(\overline{t}, \overline{\boldsymbol{x}}) = \tilde{H}(t, \boldsymbol{x}).$$

The saturation and the relative conductivity functions do not depend on the scale change and are thus assumed to be of order one. Of course,

$$\theta(L_z\overline{P}) = \theta(P), \qquad \kappa(L_z\overline{P}) = \kappa(P). \tag{A.41}$$

For the conductivity tensors, we set

$$\overline{\boldsymbol{K}}_0(\overline{\boldsymbol{x}}, \overline{z}) = \boldsymbol{K}_0(\boldsymbol{x}, z), \quad \overline{\boldsymbol{S}}_0(\overline{\boldsymbol{x}}, \overline{z}) = \boldsymbol{S}_0(\boldsymbol{x}, z), \quad \overline{\boldsymbol{N}}_0(\overline{\boldsymbol{x}}, \overline{z}) = \boldsymbol{N}_0(\boldsymbol{x}, z), \quad \overline{\boldsymbol{B}}_0(\overline{\boldsymbol{x}}, \overline{z}) = \boldsymbol{B}_0(\boldsymbol{x}, z), \tag{A.42}$$

$$\overline{\boldsymbol{K}}(\overline{H})(\overline{t}, \overline{\boldsymbol{x}}) = L_z \int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \kappa\big(\rho\, g(\overline{H}(\overline{t}, \overline{\boldsymbol{x}}) - \overline{z})\big)\overline{\boldsymbol{S}}_0\, d\overline{z}. \tag{A.43}$$

$$\overline{\boldsymbol{J}}(\overline{H})(\overline{t}, \overline{\boldsymbol{x}}) = \overline{\boldsymbol{K}}(\overline{H})(\overline{t}, \overline{\boldsymbol{x}}) - L_z \int_{\overline{h}_0(\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \kappa\big(\rho\, g(\overline{H}(\overline{t}, \overline{\boldsymbol{x}}) - \overline{z})\big)\overline{\boldsymbol{N}}_0\, d\overline{z}. \tag{A.44}$$

We choose (A.42) to simplify the presentation. Indeed, we could also introduce reference quantities $\boldsymbol{K}$ and $\boldsymbol{M}$ such that $\boldsymbol{K}\overline{\boldsymbol{K}}_0(\overline{\boldsymbol{x}}, \overline{z}) = \boldsymbol{K}_0(\boldsymbol{x}, z)$ and $\boldsymbol{M}\overline{\boldsymbol{M}}_0(\overline{\boldsymbol{x}}, \overline{z}) = \boldsymbol{M}_0(\boldsymbol{x}, z)$, and then carry out the same study assuming that $\boldsymbol{K}/L_x = \mathcal{O}(\varepsilon)$ and $\boldsymbol{M}/L_x = \mathcal{O}(\varepsilon)$. Finally, the rescaled source term is defined by

$$\overline{F}(\overline{t}, \overline{\boldsymbol{x}}) = F(t, \boldsymbol{x}).$$

Dimensionless compressible Richards problem. As the aquifer is assumed to be thin with respect to its horizontal width, the quantity L_z/L_x is accordingly small. We choose to consider an aquifer with a fixed height of order $L_z = 1$ and a large horizontal dimension $L_x = 1/\varepsilon$ for $\varepsilon \ll 1$. Then, the mass conservation equation, which depends on the choice of the time scaling T,

$$\frac{\overline{T}}{T}\partial_{\overline{t}}\theta(\overline{P}) + \alpha_p \frac{\overline{T}}{T} s(\overline{P})\partial_{\overline{t}}\overline{P} + \varepsilon\nabla_{\overline{\boldsymbol{x}}} \cdot (\overline{\boldsymbol{v}}) + \partial_{\overline{z}}\overline{\boldsymbol{v}} \cdot \boldsymbol{e}_3 = 0 \qquad \text{in }]0, \overline{T}[\times\overline{\Omega} \tag{A.45}$$

is associated with the following Darcy's law and boundary conditions:

$$\begin{cases} \overline{\boldsymbol{v}} = -\kappa(\overline{P})\,\overline{\boldsymbol{K}}_0\Big(\dfrac{\varepsilon}{\rho g}\nabla_{\overline{x}}\overline{P} + \Big(\dfrac{1}{\rho g}\partial_{\overline{z}}\overline{P}+1\Big)\boldsymbol{e}_3\Big) & \text{in }]0,\overline{T}[\times\overline{\Omega}, \\ \alpha\,\overline{P}\left(\varepsilon^2\,\|\nabla_{\overline{x}}\overline{h}_{\rm soil}\|^2+1\right)^{1/2} + \beta\,\overline{\boldsymbol{v}}\cdot\left(\boldsymbol{e}_3-\varepsilon\,\nabla_{\overline{x}}\overline{h}_{\rm soil}\right) = \left(\varepsilon^2\,\|\nabla_{\overline{x}}\overline{h}_{\rm soil}\|^2+1\right)^{1/2}\overline{F} & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\rm soil}, \\ \overline{\boldsymbol{v}}\cdot\overline{\boldsymbol{n}} = 0 & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\rm ver}, \\ \overline{\boldsymbol{v}}\cdot\left(\varepsilon\nabla_{\overline{x}}\overline{h}_{\rm bot}-\boldsymbol{e}_3\right) = 0 & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\rm bot}. \end{cases} \tag{A.46}$$

Dimensionless coupled model. For the same parameter $\varepsilon \ll 1$, the rescaled version of the coupled problem (2.2)–(2.8) is given by the 1d-Richards equation in the transition zone

$$\begin{cases} \dfrac{\overline{T}}{T}\partial_{\overline{t}}\theta(\overline{P}) + \alpha_p\dfrac{\overline{T}}{T}s(\overline{P})\partial_{\overline{t}}\overline{P} + \partial_{\overline{z}}\big(\overline{\boldsymbol{q}}\cdot\boldsymbol{e}_3\big) + \varepsilon\nabla_{\overline{x}}\cdot\overline{\boldsymbol{q}} = 0 & \text{for } \overline{t}\in]0,\overline{T}[\,, \quad (\overline{\boldsymbol{x}},\overline{z})\in\Omega_{\overline{h}}^{+}(\overline{t}), \\ \alpha\,\overline{P}\left(\varepsilon^2\,\|\nabla'\overline{h}_{\rm soil}\|^2+1\right)^{1/2} + \beta\,\overline{\boldsymbol{q}}\cdot\left(\boldsymbol{e}_3-\varepsilon\,\nabla'\overline{h}_{\rm soil}\right) = \left(\varepsilon^2\,\|\nabla'\overline{h}_{\rm soil}\|^2+1\right)^{1/2}\overline{F} & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\rm soil}, \\ \boldsymbol{q}\cdot\boldsymbol{n} = 0 & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\rm ver} \\ \overline{P}\big(\overline{t},\overline{\boldsymbol{x}},\overline{h}(\overline{t},\overline{\boldsymbol{x}})\big) = \rho\,g\big(\overline{H}(\overline{t},\overline{\boldsymbol{x}})-\overline{h}(\overline{t},\overline{\boldsymbol{x}})\big) & \text{for } (\overline{t},\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x, \\ \overline{P}(0,\overline{\boldsymbol{x}},\overline{z}) = \overline{P_{\rm init}}(\overline{\boldsymbol{x}},\overline{z}) & \text{for } (\overline{\boldsymbol{x}},\overline{z})\in\Omega_{\overline{h}}^{+}(0), \end{cases} \tag{A.47}$$

the pressure problem in the water table

$$\overline{P}(\overline{t},\overline{\boldsymbol{x}},\overline{z}) = \rho\,g\big(\overline{H}(\overline{t},\overline{\boldsymbol{x}})-\overline{z}\big) \qquad \text{for } \overline{t}\in[0,\overline{T}[\,, \quad (\overline{\boldsymbol{x}},\overline{z})\in\Omega_{\overline{h}}^{-}(\overline{t}), \tag{A.48}$$

the hydraulic head problem

$$\begin{cases} \rho\,g\,\alpha_p(\overline{h}-\overline{h}_{\rm bot})\dfrac{\overline{T}}{T}\partial_{\overline{t}}\overline{H} - \varepsilon^2\,\nabla'\cdot\big(\overline{\boldsymbol{J}}(\overline{H})\,\nabla'\overline{H}\big) = -\big(\overline{\boldsymbol{q}}\cdot(\boldsymbol{e}_3-\varepsilon\nabla\overline{h})\big)|_{\Gamma_{\overline{h}}^{+}} & \text{for } (\overline{t},\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x, \\ \overline{\boldsymbol{J}}(\overline{H})\nabla'\overline{H}\cdot\overline{\boldsymbol{n}} = 0 & \text{for } (\overline{t},\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\partial\overline{\Omega}_x, \\ \overline{H}(0,\overline{\boldsymbol{x}}) = \overline{H_{\rm init}}(\overline{\boldsymbol{x}}) & \text{for } \overline{\boldsymbol{x}}\in\overline{\Omega}_x, \end{cases} \tag{A.49}$$

where the first equation of (A.49) also reads

$$\begin{aligned} -\varepsilon^2\,\nabla'\cdot\Big(\overline{\boldsymbol{J}}(\overline{H})\,\nabla'\overline{H}\Big) = -\big(\overline{\boldsymbol{q}}\cdot(\boldsymbol{e}_3-\varepsilon\nabla\overline{h}_{\rm soil})\big)|_{\overline{\Gamma}_{\rm soil}} \\ -\frac{\overline{T}}{T}\left(\int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h_{\rm soil}}(\overline{\boldsymbol{x}})}\phi\,\partial_{\overline{t}}s(\overline{P})+\alpha_P\,s(\overline{P})\,\partial_{\overline{t}}\overline{P}\,d\overline{z}\right) - \varepsilon\nabla'\cdot\Big(\int_{\overline{h}(\overline{t},\overline{\boldsymbol{x}})}^{\overline{h_{\rm soil}}(\overline{\boldsymbol{x}})}\overline{\boldsymbol{q}}\Big) \quad \text{in } \in]0,\overline{T}[\times\overline{\Omega}_x, \end{aligned} \tag{A.50}$$

the definition of the interface separating the two different types of flows

$$\overline{h}(\overline{t},\overline{\boldsymbol{x}}) = \max\Big\{\min\Big\{\overline{H}(\overline{t},\overline{\boldsymbol{x}})-\frac{P_s}{\rho\,g},\overline{h}_{\max}(\overline{\boldsymbol{x}})\Big\},\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})\Big\} \qquad \text{for } (\overline{t},\overline{\boldsymbol{x}})\in[0,\overline{T}[\times\overline{\Omega}_x, \tag{A.51}$$

and finally the velocity problem

$$\begin{cases} \overline{\boldsymbol{v}} = \overline{\boldsymbol{q}}+\overline{\boldsymbol{w}} & \text{for } \overline{t}\in]0,\overline{T}[\,, \ (\overline{\boldsymbol{x}},\overline{z})\in\overline{\Omega}, \\ \overline{\boldsymbol{q}} = -\kappa(\overline{P})\,\overline{K}_{zz}\Big(\dfrac{1}{\rho\,g}\partial_{\overline{z}}\overline{P}+1\Big)\boldsymbol{e}_3 - \varepsilon\dfrac{\kappa(\overline{P})}{\rho\,g}\overline{\boldsymbol{G}}_0\nabla\overline{P}, & \text{for } \overline{t}\in]0,\overline{T}[\,, \ (\overline{\boldsymbol{x}},\overline{z})\in\overline{\Omega}. \\ \overline{\boldsymbol{w}} = -\varepsilon\,\kappa\big(\rho\,g(\overline{H}-\overline{z})\big)\,\overline{\boldsymbol{A}}_0\nabla\overline{H} & \text{for } \overline{t}\in]0,\overline{T}[\,, \ (\overline{\boldsymbol{x}},\overline{z})\in\overline{\Omega}, \end{cases} \tag{A.52}$$

where

$$\boldsymbol{G}_0 = \begin{pmatrix} \boldsymbol{N}_0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Effective problems

We are interested in the asymptotic behavior of the flow, that is, of the problems (A.45)-(A.46) and (A.47)-(A.52), for small and large values of T. For the asymptotic analysis, the question is related to the behavior of the dimensionless models above. More precisely, we aim to describe the effective flow obtained for short, intermediate, and long time scales, that is, $T=\overline{T}$, $T=\varepsilon^{-1}\overline{T}$, and $T=\varepsilon^{-2}\overline{T}$, respectively.

Asymptotic expansion. We introduce the following formal asymptotics for the pressure and the velocity:

$$\overline{P}_\varepsilon^\gamma = \overline{P}_0^\gamma + \varepsilon\,\overline{P}_1^\gamma + \varepsilon^2\,\overline{P}_2^\gamma + \dots \qquad \overline{\boldsymbol{v}}_\varepsilon^\gamma = \overline{\boldsymbol{v}}_0^\gamma + \varepsilon\,\overline{\boldsymbol{v}}_1^\gamma + \varepsilon^2\,\overline{\boldsymbol{v}}_2^\gamma + \dots. \tag{A.53}$$

We emphasize that no arbitrary scaling is imposed; in particular, we do not assume, as in [12], that the vertical velocity is significantly smaller than the horizontal velocity when the ratio ϵ is small. We also assume the existence of formal asymptotics for the auxiliary unknowns in (2.2)–(2.8)

$$\begin{aligned} \overline{\boldsymbol{q}}_\varepsilon^\gamma &= \overline{\boldsymbol{q}}_0^\gamma + \varepsilon\,\overline{\boldsymbol{q}}_1^\gamma + \varepsilon^2\,\overline{\boldsymbol{q}}_2^\gamma + \dots & \overline{\boldsymbol{w}}_\varepsilon^\gamma &= \overline{\boldsymbol{w}}_0^\gamma + \varepsilon\,\overline{\boldsymbol{w}}_1^\gamma + \varepsilon^2\,\overline{\boldsymbol{w}}_2^\gamma + \dots \\ \overline{H}_\varepsilon^\gamma &= \overline{H}_0 + \varepsilon\,\overline{H}_1^\gamma + \varepsilon^2\,\overline{H}_2^\gamma + \dots & \overline{h}_\varepsilon^\gamma &= \overline{h}_0^\gamma + \varepsilon\,\overline{h}_1^\gamma + \varepsilon^2\,\overline{h}_2^\gamma + \dots, \end{aligned} \tag{A.54}$$

and for the source terms

$$\overline{F}_\varepsilon = \overline{F}_0 + \varepsilon\,\overline{F}_1 + \varepsilon^2\,\overline{F}_2 + \dots. \tag{A.55}$$

Moreover, as θ and κ are piecewise $\mathscr{C}^\infty$, we write

$$\begin{aligned} \theta(\overline{P}_\varepsilon^\gamma) &= \theta(\overline{P}_0^\gamma) + \varepsilon(\overline{P}_1^\gamma + \varepsilon\,\overline{P}_2^\gamma + \dots)\theta'(\overline{P}_0^\gamma) + \frac{\varepsilon^2}{2}(\overline{P}_1^\gamma + \varepsilon\,\overline{P}_2^\gamma + \dots)^2\theta''(\overline{P}_0^\gamma) + \dots \\ \kappa(\overline{P}_\varepsilon^\gamma) &= \kappa(\overline{P}_0^\gamma) + \varepsilon(\overline{P}_1^\gamma + \varepsilon\,\overline{P}_2^\gamma + \dots)\kappa'(\overline{P}_0^\gamma) + \frac{\varepsilon^2}{2}(\overline{P}_1^\gamma + \varepsilon\,P_2^\gamma + \dots)^2\kappa''(\overline{P}_0^\gamma) + \dots \end{aligned} \tag{A.56}$$

Effective problems in the main order. Let us introduce the following effective problems:

- Related to the short time scale ($T = \overline{T}$),

$$\begin{cases} \phi\,\partial_{\overline{t}} s(\overline{P}_0) - \partial_{\overline{z}}\overline{u}_0 = 0 & \text{in }]0,\overline{T}[\times\Omega \\ \overline{u}_0 = -\kappa(\overline{P}_0)\,\overline{K}_{zz}\Big(\dfrac{1}{\rho g}\partial_{\overline{z}}\overline{P}_0 + 1\Big) & \text{in }]0,\overline{T}[\times\Omega \\ \alpha\,\overline{P}_0 + \beta\,\overline{u}_0 = \overline{F}_0 & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\text{soil}} \\ \overline{u}_0 = 0 & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\text{bot}} \end{cases} \tag{A.57}$$

- Related to the non-short time scales ($T = \varepsilon^{-1}\overline{T}$ or $T = \varepsilon^{-2}\overline{T}$),

$$\begin{cases} \overline{P}_0(t,\boldsymbol{x},z) = \rho\,g\big(\overline{H}_0(t,\boldsymbol{x}) - \overline{z}\big) & \text{in }]0,\overline{T}[\times\overline{\Omega} \\ \overline{\boldsymbol{v}}_0 = 0 & \text{in }]0,\overline{T}[\times\overline{\Omega} \end{cases} \tag{A.58}$$

- related to the non-short time scales ($T = \varepsilon^{-1}\overline{T}$ or $T = \varepsilon^{-2}\overline{T}$) if $\alpha \neq 0$,

$$\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) = \frac{\overline{F}_0(\overline{t},\overline{\boldsymbol{x}})}{\alpha\,\rho\,g} + \overline{h}_{\text{soil}}(\overline{t},\overline{\boldsymbol{x}}) \qquad \text{in }]0,\overline{T}[\times\overline{\Omega}_x \tag{A.59}$$

- related to the intermediate time scale ($T = \varepsilon^{-1}\overline{T}$) if $\alpha = 0$ (and then $\beta \neq 0$),

$$\rho\,g\Big(\int_{\overline{h}_{\text{bot}}}^{\overline{h}_{\text{soil}}} \phi\,s'(\overline{P}_0) + \alpha_P\,s(\overline{P}_0)\,dz\Big)\partial_t\overline{H}_0 = -\frac{\overline{F}_1}{\beta} \qquad \text{in }]0,\overline{T}[\times\overline{\Omega}_x \tag{A.60}$$

- related to the long time scale ($T = \varepsilon^{-2}\overline{T}$) if $\alpha = 0$ (and then $\beta \neq 0$)

$$\begin{cases} \displaystyle\int_{\overline{h}_{\text{bot}}}^{\overline{h}_{\text{soil}}} \phi\,\partial_{\overline{t}} s(\overline{P}_0) + \alpha_P\,s(\overline{P}_0)\partial_t\overline{P}_0\,d\overline{z} - \nabla'\cdot\big(\overline{\boldsymbol{K}}(\overline{H}_0)\,\nabla'\overline{H}_0\big) = -\dfrac{\overline{F}_2}{\beta} & \text{in }]0,\overline{T}[\times\overline{\Omega}_x \\ \overline{\boldsymbol{K}}(\overline{H}_0)\,\nabla_{\overline{x}}\overline{H}_0\cdot\overline{\boldsymbol{n}} = 0 & \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\text{ver}} \end{cases} \tag{A.61}$$

  and concerning the first order of the velocity

$$\overline{\boldsymbol{v}}_1 = -\overline{\kappa}(\overline{P}_0)\,\overline{\boldsymbol{M}}_0\,\nabla_{\overline{x}}\overline{H}_0 \qquad \text{in }]0,\overline{T}[\times\overline{\Omega} \tag{A.62}$$

Main convergence result and dominant behaviors

Proposition A.7. *Let $(\overline{P}_\varepsilon^\gamma, \overline{\boldsymbol{v}}_\varepsilon^\gamma)$ be the solution to the rescaled 3d-Richards problem (A.45)–(A.46) or of the rescaled coupled model (A.47)–(A.51) for $T = \varepsilon^{-\gamma}\overline{T}$ and $\gamma \in \{0,1,2\}$. We assume that (A.53)–(A.56) hold true. The main-order terms of the pressure and the velocity of the fluid are characterized by*

(i) *$\overline{P}_0^0$ satisfies (A.57) under the additionnal assumption $\alpha_p = 0$.*

(ii) *$(\overline{P}_0^1, \overline{\boldsymbol{v}}_0^1)$ satisfies (A.58) and (A.59) if $\alpha \neq 0$, or (A.58) and (A.60) with the compatibility condition $\overline{F}_0 = 0$ if $\alpha = 0$.*

(iii) *$(\overline{P}_0^2, \overline{\boldsymbol{v}}_0^2)$ satisfies (A.58) and (A.59) if $\alpha \neq 0$, or (A.58) and (A.61) with the compatibility condition $\overline{F}_0 = \overline{F}_1 = 0$ if $\alpha = 0$. Moreover, the next-order term of the velocity $\overline{\boldsymbol{v}}_1^2$ satisfies (A.62) if $\alpha = 0$.*

The proof of this crucial proposition is postponed to the Annex A.2, and we continue with some comments related to its implications.

In fact, Proposition A.7 yields two important results: First, the dominant behaviors of the flow in shallow aquifers are characterized. More precisely, we obtain different types of flow depending on the considered time scale. The second result is that, regardless of the considered time scale, the model (2.2)–(2.8) and the compressible Richards model exhibit exactly the same dominant behaviors (with the restriction $\alpha_p = 0$ for the short time-scale). Hence, whatever the choice of G_0 and $\alpha_p \geq 0$ the model (2.2)–(2.8) is a proper approximation of the compressible 3d-Richards model when the ratio deepness/horizontal length of the aquifer is small (with the restriction $\alpha_p = 0$ in the short time-scale).

Proof of Proposition A.7 for the Richards model

The proof of Proposition A.7 consists in substituting the formal asymptotic expansion (A.53)–(A.56) into the rescaled compressible 3d-Richards problem (A.45)–(A.46). A cascade of equations follows by identifying the powers of ε. Then, we characterize the main-order terms in the expansion (A.53). For simplicity, we will suppress the superscript γ in the unknowns.

General relations. We first state the relations holding on every time scale (*i.e.*, for all $\gamma \in \{0,1,2\}$). By substituting the asymptotic expansion (A.53) into the first equation in (A.46), we obtain the following relations holding in $]0,\overline{T}[\times\Omega$:

$$\begin{cases} \overline{\boldsymbol{v}}_0 = -\kappa(\overline{P}_0)\left(\dfrac{1}{\rho\, g}\partial_{\overline{z}}\overline{P}_0 + 1\right)\overline{\boldsymbol{K}}_0\,\boldsymbol{e}_3, \\ \overline{\boldsymbol{v}}_1 = -\dfrac{\kappa(\overline{P}_0)}{\rho\, g}\,\overline{\boldsymbol{K}}_0\left(\nabla_{\overline{x}}\overline{P}_0 + \partial_{\overline{z}}\overline{P}_1\,\boldsymbol{e}_3\right) - \kappa'(\overline{P}_0)\,\overline{P}_1\left(\dfrac{1}{\rho\, g}\partial_{\overline{z}}\overline{P}_0 + 1\right)\overline{\boldsymbol{K}}_0\,\boldsymbol{e}_3. \end{cases} \tag{A.63}$$

The same process in the three last equations in (A.46) yields the following relations in $]0,\overline{T}[$:

- on $\overline{\Gamma}_{\mathrm{soil}}$

$$\begin{cases} \alpha\,\overline{P}_0 + \beta\,\overline{\boldsymbol{v}}_0\cdot\boldsymbol{e}_3 = \overline{F}_0, \qquad \alpha\,\overline{P}_1 + \beta\left(\overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 - \overline{\boldsymbol{v}}_0\cdot\nabla_{\overline{x}}\overline{h}_{\mathrm{soil}}\right) = \overline{F}_1, \\ \alpha\left(\overline{P}_2 + \dfrac{1}{2}\|\nabla_{\overline{x}}\overline{h}_{\mathrm{soil}}\|^2\,\overline{P}_0\right) + \beta\left(\overline{\boldsymbol{v}}_2\cdot\boldsymbol{e}_3 - \overline{\boldsymbol{v}}_1\cdot\nabla_{\overline{x}}\overline{h}_{\mathrm{soil}}\right) = \dfrac{1}{2}\|\nabla_{\overline{x}}\overline{h}_{\mathrm{soil}}\|^2\,\overline{F}_0 + \overline{F}_2; \end{cases} \tag{A.64}$$

- on $\overline{\Gamma}_{\mathrm{bot}}$, for all $k\in\mathbb{N}^*$

$$\overline{\boldsymbol{v}}_0\cdot\boldsymbol{e}_3 = 0, \qquad \overline{\boldsymbol{v}}_{k-1}\cdot\nabla_{\overline{x}}\overline{h}_{\mathrm{bot}} = \overline{\boldsymbol{v}}_k\cdot\boldsymbol{e}_3; \tag{A.65}$$

- on $\overline{\Gamma}_{\mathrm{ver}}$, for all $k\in\mathbb{N}$

$$\overline{\boldsymbol{v}}_k\cdot\overline{\boldsymbol{n}} = 0. \tag{A.66}$$

Short time case. We prove the first claim of Proposition A.7, which is associated with the short characteristic time scale $T = \varepsilon^{-\gamma}\overline{T}$ for $\gamma = 0$. We assume then that $\alpha_p = 0$. The equation (A.45) here reads

$$\partial_{\overline{t}}\theta(\overline{P}) + \varepsilon\nabla_{\overline{\boldsymbol{x}}}\cdot(\overline{\boldsymbol{v}}) + \partial_{\overline{z}}\overline{\boldsymbol{v}}\cdot\boldsymbol{e}_3 = 0.$$

Some computations show that the main-order part of the last equation is

$$\partial_{\overline{t}}\theta(\overline{P}_0) + \partial_{\overline{z}}\overline{\boldsymbol{v}}_0\cdot\boldsymbol{e}_3 = 0 \qquad \text{in }]0,\overline{T}[\times\overline{\Omega}.$$

This equation combined with the first equations in (A.63), (A.64), and (A.65) yield exactly the system (A.57) (since $\overline{\boldsymbol{v}}_0\cdot\boldsymbol{e}_3 = \overline{u}_0$). The first claim of Proposition A.7 is thus proved.

Intermediate time case. In this part, we prove the second claim of Proposition A.7, which is associated with the intermediate time scale $T = \varepsilon^{-\gamma}\overline{T}$ for $\gamma = 1$. Equation (A.45) is now

$$\varepsilon\phi\,\partial_{\overline{t}}s(\overline{P}) + \varepsilon\alpha_p\,s(\overline{P})\partial_{\overline{t}}\overline{P} + \varepsilon\nabla_{\overline{\boldsymbol{x}}}\cdot(\overline{\boldsymbol{v}}) + \partial_{\overline{z}}\overline{\boldsymbol{v}}\cdot\boldsymbol{e}_3 = 0. \tag{A.67}$$

We introduce the asymptotic expansion (A.53) into the previous equation, and we identify the main order terms. We obtain

$$\partial_{\overline{z}}\overline{\boldsymbol{v}}_0\cdot\boldsymbol{e}_3 = 0 \qquad \text{on }]0,\overline{T}[\times\overline{\Omega}.$$

Thus, $\boldsymbol{v}_0\cdot\boldsymbol{e}_3$ does not depend on $\overline{z}$. It is in fact zero by (A.65). Moreover, by the first equation in (A.63), as κ and $\overline{K}_{zz}$ are non-vanishing ($\overline{\boldsymbol{K}}_0$ is positive definite), we obtain in $]0,\overline{T}[\times\overline{\Omega}$

$$\partial_{\overline{z}}\overline{P}_0 + \rho\, g = 0 \qquad \text{and} \qquad \overline{\boldsymbol{v}}_0 = 0. \tag{A.68}$$

The existence of $\overline{H}_0 = \overline{H}_0(\overline{t},\overline{\boldsymbol{x}})$ such that

$$\overline{P}_0(\overline{t},\overline{\boldsymbol{x}},\overline{z}) = \rho\, g\big(\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) - \overline{z}\big) \qquad \text{in }]0,\overline{T}[\times\overline{\Omega} \tag{A.69}$$

follows. With (A.68)–(A.69), we have proved (A.58). Subsequently, as $\overline{\boldsymbol{v}}_0 = 0$, the first equation in (A.64) reduces to

$$\alpha\,\overline{P}_0 = \overline{F}_0 \qquad \text{on } \overline{\Gamma}_{\text{soil}}. \tag{A.70}$$

We now differentiate the computations depending on whether $\alpha = 0$.

If $\alpha \neq 0$, then for all $(\overline{t},\overline{\boldsymbol{x}}) \in]0,\overline{T}[\times\overline{\Omega}_x$, we have $\overline{P}_0(\overline{t},\overline{\boldsymbol{x}},\overline{h}_{\text{soil}}(\overline{t},\overline{\boldsymbol{x}})) = \overline{F}_0(\overline{t},\overline{\boldsymbol{x}})/\alpha$. Accordingly, by (A.69),

$$\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) = \frac{\overline{F}_0(\overline{t},\overline{\boldsymbol{x}})}{\alpha\,\rho\,g} + \overline{h}_{\text{soil}}(\overline{t},\overline{\boldsymbol{x}}).$$

This completes the proof of the second claim of Proposition A.7 in the case $\alpha \neq 0$.

If $\alpha = 0$ (then $\beta \neq 0$), Equation (A.70) only implies the compatibility condition $\overline{F}_0 = 0$. We should exploit the next-order terms in the asymptotic expansion to conclude the analysis of the effective problem. Identifying the coefficients associated with ε^1 in Equation (A.67), we obtain

$$\partial_{\overline{t}}\theta(\overline{P}_0) + \alpha_p\, s(\overline{P}_0)\partial_{\overline{t}}\overline{P}_0 + \partial_{\overline{z}}\overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 = 0 \qquad \text{in }]0,\overline{T}[\times\overline{\Omega}.$$

To eliminate $\overline{\boldsymbol{v}}_1$, we integrate vertically the last equation on $]\overline{h}_{\text{bot}},\overline{h}_{\text{soil}}[$. As (A.69) implies that $\partial_{\overline{t}}(s(\overline{P}_0)) = \rho\,g\,s'(\overline{P}_0)\,\partial_{\overline{t}}\overline{H}_0$, we write

$$\rho\,g\Big(\int_{\overline{h}_{\text{bot}}}^{\overline{h}_{\text{soil}}} \theta'(\overline{P}_0) + \alpha_p\, s(\overline{P}_0)\,d\overline{z}\Big)\partial_{\overline{t}}\overline{H}_0 + (\overline{\boldsymbol{v}}_1|_{\overline{h}_{\text{soil}}} - \overline{\boldsymbol{v}}_1|_{\overline{h}_{\text{bot}}})\cdot\boldsymbol{e}_3 = 0. \tag{A.71}$$

From the second equations in (A.64) and (A.65) in the case where $\alpha = 0$ and $\overline{\boldsymbol{v}}_0 = 0$, it follows that

$$\overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 = \overline{F}_1/\beta \quad \text{on } \overline{\Gamma}_{\text{soil}} \qquad \text{and} \qquad \overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 = 0 \quad \text{on } \overline{\Gamma}_{\text{bot}}.$$

Accordingly, Equation (A.71) becomes

$$\rho\,g\Big(\int_{\overline{h}_{\text{bot}}}^{\overline{h}_{\text{soil}}} \theta'(\overline{P}_0) + \alpha_p\, s(\overline{P}_0)\,d\overline{z}\Big)\partial_{\overline{t}}\overline{H}_0 = -\frac{\overline{F}_1}{\beta}. \tag{A.72}$$

Finally, collecting Equations (A.69) and (A.72), we obtain $\overline{\boldsymbol{v}}_0 = 0$ and

$$\begin{cases} \overline{P}_0(\overline{t},\overline{\boldsymbol{x}},\overline{z}) = \rho\,g\big(\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) - \overline{z}\big) & \text{in }]0,\overline{T}[\times\overline{\Omega} \\ \rho\,g\Big(\displaystyle\int_{\overline{h}_{\text{bot}}}^{\overline{h}_{\text{soil}}} \theta'(\overline{P}_0) + \alpha_p\, s(\overline{P}_0)\,d\overline{z}\Big)\partial_{\overline{t}}\overline{H}_0 = -\frac{\overline{F}_1}{\beta} & \text{in }]0,\overline{T}[\times\overline{\Omega}_x \end{cases}$$

which correspond to the second claim of Proposition A.7 in the case $\alpha = 0$.

Long time case. In this part, we prove the third claim of Proposition A.7, which is associated with the intermediate time scale $T = \varepsilon^{-\gamma}\overline{T}$ for $\gamma = 2$. Equation (A.45) now reads

$$\varepsilon^2\partial_{\overline{t}}\theta(\overline{P}) + \varepsilon^2\alpha_p\, s(\overline{P})\partial_{\overline{t}}\overline{P} + \varepsilon\nabla_{\overline{\boldsymbol{x}}}\cdot(\overline{\boldsymbol{v}}) + \partial_{\overline{z}}\overline{\boldsymbol{v}}\cdot\boldsymbol{e}_3 = 0. \tag{A.73}$$

We substitute the asymptotic expansion (A.53) into the previous equation. The main-order part of the equation is $\partial_z(\overline{\boldsymbol{v}}_0\cdot\boldsymbol{e}_3) = 0$, which, as before, leads to (A.58) for some function $\overline{H}_0$ that does not depend on $\overline{z}$. The same relation (A.70) holds, and the characterization of $\overline{H}_0$ depends on the values of α. As before, if $\alpha \neq 0$, we have (A.59).

It remains to consider the case $\alpha = 0$ and to exhibit the equations of system (A.61). In this case, the compatibility condition $\overline{F}_0 = 0$ is necessary as before owing to (A.70). The characterization of $\overline{H}_0$ requires the next-order part of Equation (A.73), namely,

$$0 = \nabla_{\overline{\boldsymbol{x}}}\cdot\overline{\boldsymbol{v}}_0 + \partial_{\overline{z}}\overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 = \partial_{\overline{z}}\overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 \tag{A.74}$$

where the second equality holds because $\overline{\boldsymbol{v}}_0 = 0$. Moreover, the second equations in (A.64) and (A.65) for $k = 1$ lead to (as $\alpha = 0$)

$$\beta\,\overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 = \overline{F}_1 \quad \text{on } \overline{\Gamma}_{\text{soil}} \qquad \text{and} \qquad \overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3 = 0 \quad \text{on } \overline{\Gamma}_{\text{bot}}. \tag{A.75}$$

Then, the vertical component $\overline{\boldsymbol{v}}_1\cdot\boldsymbol{e}_3$ of the velocity (which is constant by (A.74)) is zero. Moreover, the second compatibility condition $\overline{F}_1 = 0$ appears according to (A.75). Using the second equation in (A.63) and bearing in mind that $(\rho\,g)^{-1}\partial_{\overline{z}}\overline{P}_0 + 1 = 0$, we obtain

$$\overline{\boldsymbol{v}}_1 = -\frac{\kappa(\overline{P}_0)}{\rho\,g}\,\overline{\boldsymbol{K}}_0\Big(\nabla_{\overline{\boldsymbol{x}}}\overline{P}_0 + \partial_{\overline{z}}\overline{P}_1\,\boldsymbol{e}_3\Big).$$

As $\overline{\boldsymbol{v}}_1 \cdot \boldsymbol{e}_3 = 0$, using the same notation for $\overline{\boldsymbol{K}}_0$ as in (1.5), we compute $\partial_{\overline{z}} \overline{P}_1$:

$$\partial_{\overline{z}} \overline{P}_1 = -\frac{1}{\overline{K}_{zz}} \overline{\boldsymbol{K}}_0 \nabla_{\overline{x}} \overline{P}_0 \cdot \boldsymbol{e}_3 = 0.$$

Subsequently, using the relation $\overline{P}_0 = \rho\, g(\overline{H}_0 - \overline{z})$ in the last equation, we obtain

$$\overline{\boldsymbol{v}}_1 = -\kappa(\overline{P}_0)\, \overline{\boldsymbol{M}}_0 \nabla_{\overline{x}} \overline{H}_0 \quad \text{with} \quad \overline{\boldsymbol{M}}_0 = \begin{pmatrix} \boldsymbol{K}_{xx} & 0 \\ 0 & 0 \end{pmatrix}. \tag{A.76}$$

Equation (A.66) for $k = 1$ now leads to $\overline{\boldsymbol{v}}_1 \cdot \overline{\boldsymbol{n}} = 0$ on $\overline{\Gamma}_{\mathrm{ver}}$. As $\kappa(\overline{P}_0)$ does not vanish, we obtain the last equation in (A.61). After identifying the coefficients associated with ε^2 in Equation (A.73), we obtain

$$\partial_{\overline{t}} \theta(\overline{P}_0) + \alpha_p s(\overline{P}_0) \partial_{\overline{t}} \overline{P}_0 + \nabla_{\overline{\boldsymbol{x}}} \cdot (\overline{\boldsymbol{v}}_1) + \partial_{\overline{z}} \overline{\boldsymbol{v}}_2 \cdot \boldsymbol{e}_3 = 0. \tag{A.77}$$

By (A.58), (A.76), and the fact that $\alpha = F_0 = 0$, the third equation in (A.64) and the second equation in (A.65) for $k = 2$ become

$$\overline{\boldsymbol{v}}_2 \cdot \boldsymbol{e}_3 - \overline{\boldsymbol{v}}_1 \cdot \nabla_{\overline{x}} \overline{h}_{\mathrm{soil}} = \overline{F}_2/\beta \quad \text{on } \overline{\Gamma}_{\mathrm{soil}}, \qquad \overline{\boldsymbol{v}}_2 \cdot \boldsymbol{e}_3 - \overline{\boldsymbol{v}}_1 \cdot \nabla_{\overline{x}} \overline{h}_{\mathrm{bot}} = 0 \quad \text{on } \overline{\Gamma}_{\mathrm{bot}}. \tag{A.78}$$

To eliminate v_2 in system (A.77)–(A.78), we integrate (A.77) with respect to $\overline{z}$ on $[\overline{h}_{\mathrm{bot}}, \overline{h}_{\mathrm{soil}}]$. By the boundary conditions on $\overline{\Gamma}_{\mathrm{bot}}$ and $\overline{\Gamma}_{\mathrm{soil}}$, we obtain

$$\partial_{\overline{t}} \int_{\overline{h}_{\mathrm{bot}}}^{\overline{h}_{\mathrm{soil}}} \theta(\overline{P}_0)\, d\overline{z} + \int_{\overline{h}_{\mathrm{bot}}}^{\overline{h}_{\mathrm{soil}}} \alpha_p s(\overline{P}_0) \partial_{\overline{t}} \overline{P}_0\, d\overline{z} + \int_{\overline{h}_{\mathrm{bot}}}^{\overline{h}_{\mathrm{soil}}} \nabla_{\overline{\boldsymbol{x}}} \cdot \overline{\boldsymbol{v}}_1\, d\overline{z}$$
$$+ \overline{\boldsymbol{v}}_1|_{\overline{h}_{\mathrm{soil}}} \cdot \nabla_{\overline{x}} \overline{h}_{\mathrm{soil}} + \frac{\overline{F}_2}{\beta} - \overline{\boldsymbol{v}}_1|_{\overline{h}_{\mathrm{bot}}} \cdot \nabla_{\overline{x}} \overline{h}_{\mathrm{bot}} = 0.$$

We use the Leibniz rule in the second integral and obtain

$$\int_{\overline{h}_{\mathrm{bot}}}^{\overline{h}_{\mathrm{soil}}} \partial_{\overline{t}} \theta(\overline{P}_0) + \alpha_p\, s(\overline{P}) \partial_{\overline{t}} \overline{P}_0\, d\overline{z} + \nabla_{\overline{\boldsymbol{x}}} \cdot \Big(\int_{\overline{h}_{\mathrm{bot}}}^{\overline{h}_{\mathrm{soil}}} \overline{\boldsymbol{v}}_1\, d\overline{z} \Big) = -\frac{\overline{F}_2}{\beta}. \tag{A.79}$$

Using the first equation in (A.76) and the averaged conductivity $\overline{\boldsymbol{K}}$ defined in (A.43), we obtain

$$\int_{\overline{h}_{\mathrm{bot}}}^{\overline{h}_{\mathrm{soil}}} \overline{\boldsymbol{v}}_1\, d\overline{z} = -\int_{\overline{h}_{\mathrm{bot}}}^{\overline{h}_{\mathrm{soil}}} \kappa(\overline{P}_0)\, \overline{\boldsymbol{M}}_0 \nabla_{\overline{x}} \overline{H}_0\, d\overline{z} = -\overline{\boldsymbol{K}}(\overline{H}_0) \nabla_{\overline{x}} \overline{H}_0.$$

The last equation associated with Equation (A.79) is exactly the system (A.61). This completes the proof of the last claim of Proposition (A.7).

Proof of Proposition A.7 for the coupled models

The strategy of the proof is exactly the same as that in the previous subsection.

General relations. Let $\gamma \in \{0, 1, 2\}$. Using the expansion (A.53)–(A.56), we identify the powers of ε in all the equations in (A.47)–(A.52) that do not depend on the time scale T. We obtain from the second equation in (A.52)

$$\begin{cases} \overline{\boldsymbol{q}}_0 = -\kappa(\overline{P}_0)\, \overline{K}_{zz} \Big(\dfrac{1}{\rho\, g} \partial_{\overline{z}} \overline{P}_0 + 1 \Big) \boldsymbol{e}_3 & \text{in }]0, \overline{T}[\times \overline{\Omega}, \\ \overline{\boldsymbol{q}}_1 = -\overline{K}_{zz} \dfrac{\kappa(\overline{P}_0)}{\rho\, g} \partial_{\overline{z}} \overline{P}_1 \boldsymbol{e}_3 - \kappa'(\overline{P}_0) \overline{K}_{zz} \overline{P}_1 \Big(\dfrac{1}{\rho\, g} \partial_{\overline{z}} \overline{P}_0 + 1 \Big) \boldsymbol{e}_3 - \dfrac{\kappa(\overline{P}_0)}{\rho\, g} \boldsymbol{G}_0 \nabla_{\overline{x}} \overline{P}_0 & \text{in }]0, \overline{T}[\times \overline{\Omega}, \end{cases} \tag{A.80}$$

from the third equation in (A.52)

$$\overline{\boldsymbol{w}}_0 = 0, \qquad \overline{\boldsymbol{w}}_1 = -\kappa\big(\rho\, g(\overline{H}_0 - \overline{z})\big) \overline{\boldsymbol{A}}_0 \nabla_{\overline{x}} \overline{H}_0 \qquad \text{in }]0, \overline{T}[\times \overline{\Omega} \tag{A.81}$$

and from the first equation in (A.52)

$$\begin{cases} \overline{\boldsymbol{v}}_0 = \overline{\boldsymbol{q}}_0 + \overline{\boldsymbol{w}}_0 = \overline{\boldsymbol{q}}_0 = -\kappa(\overline{P}_0) \overline{K}_{zz} \Big(\dfrac{1}{\rho\, g} \partial_{\overline{z}} \overline{P}_0 + 1 \Big) \boldsymbol{e}_3 & \text{in }]0, \overline{T}[\times \overline{\Omega}, \\ \overline{\boldsymbol{v}}_1 = \overline{\boldsymbol{q}}_1 + \overline{\boldsymbol{w}}_1 & \text{in }]0, \overline{T}[\times \overline{\Omega}. \end{cases} \tag{A.82}$$

It follows from (A.48) that for $\overline{t} \in]0, \overline{T}[$ and $(\overline{\boldsymbol{x}}, \overline{z}) \in \Omega^{-}_{\overline{h}_0}(\overline{t})$

$$\overline{P}_0(\overline{t}, \overline{\boldsymbol{x}}, \overline{z}) = \rho\, g\big(\overline{H}_0(\overline{t}, \overline{\boldsymbol{x}}) - \overline{z}\big), \quad \overline{P}_k(\overline{t}, \overline{\boldsymbol{x}}, \overline{z}) = \rho\, g \overline{H}_k(\overline{t}, \overline{\boldsymbol{x}})\ \forall k > 0. \tag{A.83}$$

Equation (A.51) yields

$$\overline{h}_0(\overline{t},\overline{\boldsymbol{x}}) = \max\Big\{\min\Big\{\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) - \frac{P_s}{\rho\, g}, \overline{h}_{\max}(\overline{\boldsymbol{x}})\Big\}, \overline{h}_{\rm bot}(\overline{\boldsymbol{x}})\Big\} \qquad \text{for } (\overline{t},\overline{\boldsymbol{x}}) \in [0,\overline{T}[\times\overline{\Omega}_x \tag{A.84}$$

For the boundary conditions, we infer from the second and third equations in (A.47) and from the second equation in (A.49) that for all $k \in \mathbb{N}$,

$$\begin{cases} \alpha\,\overline{P}_0 + \beta\,\overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3 = \overline{F}_0, \qquad \alpha\,\overline{P}_1 + \beta\,(\overline{\boldsymbol{q}}_1\cdot\boldsymbol{e}_3 - \overline{\boldsymbol{q}}_0\cdot\nabla_{\overline{x}}\overline{h}_{\rm soil}) = \overline{F}_1, \\ \alpha\left(\overline{P}_2 + \dfrac{1}{2}\|\nabla_{\overline{x}}\overline{h}_{\rm soil}\|^2\,\overline{P}_0\right) + \beta\,(\overline{\boldsymbol{q}}_2\cdot\boldsymbol{e}_3 - \overline{\boldsymbol{q}}_1\cdot\nabla_{\overline{x}}\overline{h}_{\rm soil}) = \dfrac{1}{2}\|\nabla_{\overline{x}}\overline{h}_{\rm soil}\|^2\,\overline{F}_0 + \overline{F}_2 \\ \overline{P}_0\big(\overline{t},\overline{\boldsymbol{x}},\overline{h}_0(\overline{t},\overline{\boldsymbol{x}})\big) = \rho\, g\big(\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) - \overline{h}_0(\overline{t},\overline{\boldsymbol{x}})\big) \qquad \text{for } \overline{t}\in]0,\overline{T}[, \quad \overline{\boldsymbol{x}}\in\Gamma_{\overline{h}}(\overline{t}), \\ \overline{\boldsymbol{J}}(\overline{H}_0)\nabla_{\overline{x}}\overline{H}_0\cdot\overline{n} = 0 \qquad \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\rm ver}. \end{cases} \tag{A.85}$$

By (A.83) for $k=1$, $\partial_{\overline{z}}\overline{P}_1 = 0$ on $\Omega^-_{\overline{h}_0}(\overline{t})$. Then, by (A.80) and the first equation in (A.83)

$$\overline{\boldsymbol{q}}_1 = -\boldsymbol{G}_0\nabla_{\overline{x}}\overline{H}_0 \qquad \text{in } \Omega^-_{\overline{h}_0}(\overline{t}). \tag{A.86}$$

Short time case. In this part, $T=\overline{T}$, that is $\gamma=0$. We also assume here that $\alpha_p = 0$. The first equations in (A.47) and (A.49) become

$$\begin{cases} \partial_{\overline{t}}\theta(\overline{P}) + \partial_{\overline{z}}\big(\overline{\boldsymbol{q}}\cdot\boldsymbol{e}_3\big) - \varepsilon\nabla_{\overline{x}}\cdot\overline{\boldsymbol{q}} = 0 \qquad \text{for } \overline{t}\in]0,\overline{T}[\,, \quad (\overline{\boldsymbol{x}},\overline{z})\in\Omega^+_{\overline{h}}(\overline{t}), \\ -\varepsilon^2\nabla'\cdot\big(\overline{\boldsymbol{J}}(\overline{H})\,\nabla'\overline{H}\big) = -\big(\overline{\boldsymbol{q}}\cdot(\boldsymbol{e}_3 - \varepsilon\nabla\overline{h})\big)|_{\Gamma_{\overline{h}}} \quad \text{for } (\overline{t},\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x. \end{cases}$$

We identify the main-order terms appearing when (A.53)–(A.56) are substituted into the previous equations:

$$\partial_{\overline{t}}\theta(\overline{P}_0) + \partial_{\overline{z}}\big(\overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3\big) = 0 \qquad \text{for } \overline{t}\in]0,\overline{T}[\text{ and } (\overline{\boldsymbol{x}},\overline{z})\in\Omega^+_{\overline{h}_0}(\overline{t}), \tag{A.87}$$

$$0 = (\overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3)|_{\Gamma_{\overline{h}_0}} \qquad \text{for } (\overline{t},\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x. \tag{A.88}$$

From (A.82) and (A.83), we also compute $\overline{\boldsymbol{q}}_0 = 0$ in $\Omega^-_{\overline{h}_0}(\overline{t})$. In addition, from (A.84) and $\overline{R}\geq 0$, we obtain $s(\overline{P}_0) = 1$ in $\Omega^-_{\overline{h}}(\overline{t})$, so that $\overline{P}_0$ satisfies (A.87) also in $\Omega^-_{\overline{h}}(\overline{t})$ ($\overline{u}_0 = \overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3$). As the continuity of $\overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3$ is ensured by (A.88), $\overline{P}_0$ satisfies (A.87) in the entire Ω. By using (A.82), (A.85), and (A.87)-(A.88), we obtain the system (A.57), and then the first claim of Proposition A.7 holds.

Intermediate time case. In this part, $T = \varepsilon^{-1}\overline{T}$, $\gamma = 1$. The first equation in (A.47) and Equation (A.50) become

$$\begin{cases} \varepsilon\partial_{\overline{t}}\theta(\overline{P}) + \varepsilon\alpha_p\, s(\overline{P})\,\partial_{\overline{t}}\overline{P} + \partial_{\overline{z}}\big(\overline{\boldsymbol{q}}\cdot\boldsymbol{e}_3\big) + \varepsilon\nabla_{\overline{x}}\cdot\overline{\boldsymbol{q}} = 0 \qquad \text{for } \overline{t}\in]0,\overline{T}[\,, \quad (\overline{\boldsymbol{x}},\overline{z})\in\Omega^+_{\overline{h}}(\overline{t}) \\ -\varepsilon^2\nabla'\cdot\big(\overline{\boldsymbol{J}}(\overline{H})\,\nabla'\overline{H}\big) = -(\overline{\boldsymbol{q}}\cdot(\boldsymbol{e}_3 - \varepsilon\nabla_{\overline{x}}h_{\rm soil})|_{\overline{\Gamma}_{\rm soil}} \\ \qquad\qquad -\varepsilon\Big(\displaystyle\int_{\overline{h}_{\rm bot}(\overline{t},\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \partial_{\overline{t}}\theta(\overline{P}) + \alpha_p\,\theta(\overline{P})\,\partial_{\overline{t}}\overline{P}\,d\overline{z}\Big) - \varepsilon\nabla_{\overline{x}}\cdot\Big(\displaystyle\int_{\overline{h}(\overline{t},\overline{\boldsymbol{x}})}^{\overline{h_{\rm soil}}(\overline{\boldsymbol{x}})} \overline{\boldsymbol{q}}\Big) \quad \text{for } (\overline{t},\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x \end{cases} \tag{A.89}$$

The corresponding main-order relations are

$$\overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3 = 0 \qquad \text{on }]0,\overline{T}[\times\overline{\Gamma}_{\rm soil}$$

and for $\overline{t}\in]0,\overline{T}[$ and $(\overline{\boldsymbol{x}},\overline{z})\in\Omega^+_{\overline{h}}(\overline{t})$,

$$\partial_{\overline{z}}\big(\overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3\big) = 0.$$

It follows that the constant vertical component of the velocity $\overline{\boldsymbol{q}}_0\cdot\boldsymbol{e}_3$ equals zero in $\Omega^+_{\overline{h}}(\overline{t})$. We deduce from the first equation in (A.80) that the pressure $\overline{P}_0$ is affine with respect to the vertical variable, with the slope $-\rho\, g$ in $\Omega^+_{\overline{h}}(\overline{t})$. Accordingly, by the first equation in (A.83) and the continuity condition in (A.85), the first equation in (A.58) holds. In particular $\overline{\boldsymbol{q}}_0 = 0$. Using relation (A.82), we obtain the second equation in (A.58). Subsequently, by $\overline{\boldsymbol{q}}_0 = 0$ and the first equation in (A.85) for $k=0$, we obtain $\alpha\,\overline{P}_0 = \overline{F}_0$.

If $\alpha\neq 0$, then for all $(\overline{t},\overline{\boldsymbol{x}})\in]0,\overline{T}[\times\overline{\Omega}_x$, we have $P_0(\overline{t},\overline{\boldsymbol{x}},\overline{h}_{\rm soil}(\overline{t},\overline{\boldsymbol{x}})) = \overline{F}_0(\overline{t},\overline{\boldsymbol{x}})/\alpha$. According to the first equation in (A.58) (already proved in (A.83)), we have

$$\overline{H}_0(\overline{t},\overline{\boldsymbol{x}}) = \frac{\overline{F}_0(\overline{t},\overline{\boldsymbol{x}})}{\alpha\,\rho\, g} + \overline{h}_{\rm soil}(\overline{t},\overline{\boldsymbol{x}}).$$

The second claim of Proposition A.7 in the case $\alpha\neq 0$ is proved.

If $\alpha = 0$, the compatibility condition $\overline{F}_0 = 0$ is imposed by the relation $\alpha \overline{P}_0 = \overline{F}_0$. After identifying the coefficients associated with ε^1 in the second equation in (A.89), we have

$$0 = -(\overline{\boldsymbol{q}}_1 \cdot \boldsymbol{e}_3)|_{\overline{\Gamma}_{\rm soil}} - (\boldsymbol{q}_0 \cdot \nabla h_{\rm soil})|_{\overline{\Gamma}_{\rm soil}} - \Big(\int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \partial_{\overline{t}} \theta(\overline{P}_0) + \alpha_p \theta(\overline{P}_0)\, \partial_{\overline{t}} \overline{P}_0 \, d\overline{z} \Big) - \nabla_{\overline{x}} \cdot \Big(\int_{\overline{h}}^{\overline{h}_{\rm soil}} \boldsymbol{q}_0 \Big).$$

and by the first equation in (A.58) and the equality $\boldsymbol{q}_0 = 0$,

$$\rho\, g \Big(\int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \theta'(\overline{P}_0) + \alpha_p \theta(\overline{P}_0)\, d\overline{z} \Big) \partial_{\overline{t}} \overline{H}_0 = -(\overline{\boldsymbol{q}}_1 \cdot \boldsymbol{e}_3)|_{\overline{\Gamma}_{\rm soil}}.$$

As $\alpha = 0$, the first equation in (A.85) for $k = 1$ implies that $(\overline{\boldsymbol{q}}_1 \cdot \boldsymbol{e}_3)|_{\overline{\Gamma}_{\rm soil}} = \overline{F}_1 / \beta$. This completes the proof of the second claim of Proposition A.7 in the case $\alpha = 0$.

Long time case. In this part, $T = \varepsilon^{-\gamma} \overline{T}$, $\gamma = 2$. The first equation in (A.47) and Equation (A.50) are now

$$\begin{cases} \varepsilon^2 \partial_{\overline{t}} \theta(\overline{P}) + \varepsilon^2 \alpha_p \theta(\overline{P})\, \partial_{\overline{t}} \overline{P} + \partial_{\overline{z}} \big(\overline{\boldsymbol{q}} \cdot \boldsymbol{e}_3 \big) + \varepsilon \nabla_{\overline{x}} \cdot \overline{\boldsymbol{q}} = 0 \qquad \text{for } \overline{t} \in]0, \overline{T}[\,, \quad (\overline{\boldsymbol{x}}, \overline{z}) \in \Omega_{\overline{h}}^{+}(\overline{t}) \\ -\varepsilon^2 \nabla' \cdot \Big(\overline{\boldsymbol{J}}(\overline{H})\, \nabla' \overline{H} \Big) = -(\overline{\boldsymbol{q}} \cdot \boldsymbol{e}_3)|_{\overline{\Gamma}_{\rm soil}} + \varepsilon (\boldsymbol{q} \cdot \nabla \overline{h}_{\rm soil})|_{\overline{\Gamma}_{\rm soil}} \\ \qquad\qquad - \varepsilon^2 \Big(\int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \partial_{\overline{t}} \theta(\overline{P}) + \alpha_p \theta(\overline{P})\, \partial_{\overline{t}} \overline{P}\, d\overline{z} \Big) - \varepsilon \nabla_{\overline{x}} \cdot \Big(\int_{\overline{h}(\overline{t}, \overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \overline{\boldsymbol{q}} \Big) \qquad \text{for } (\overline{t}, \overline{\boldsymbol{x}}) \in]0, \overline{T}[\times \overline{\Omega}_x \end{cases} \tag{A.90}$$

As in the intermediate time case, we substitute the asymptotics (A.53)–(A.56) into the previous equations. Identifying the coefficients associated with ε^0, we obtain $\partial_{\overline{z}} (\overline{\boldsymbol{q}}_0 \cdot \boldsymbol{e}_3) = 0$ in $\Omega_{\overline{h}}^{+}(\overline{t})$ and $\overline{\boldsymbol{q}}_0 \cdot \boldsymbol{e}_3 = 0$ on $\overline{\Gamma}_{\rm soil}$. This leads to

$$\overline{\boldsymbol{q}}_0 \cdot \boldsymbol{e}_3 = 0. \tag{A.91}$$

By using the same arguments, we obtain $\overline{P}_0 = \rho\, g (\overline{H}_0 - \overline{z})$ and $\overline{\boldsymbol{q}}_0 = \overline{\boldsymbol{v}}_0 = \overline{\boldsymbol{w}}_0 = 0$ in the entire $\overline{\Omega}$. The system (A.58) is satisfied.

Identifying the coefficients associated with ε^1, we obtain $\partial_{\overline{z}} (\overline{\boldsymbol{q}}_1 \cdot \boldsymbol{e}_3) = -\nabla_{\overline{x}} \cdot \overline{\boldsymbol{q}}_0 = 0$ in $\Omega_{\overline{h}}^{+}(\overline{t})$ and $\overline{\boldsymbol{q}}_1 \cdot \boldsymbol{e}_3 = -\nabla_{\overline{x}} \cdot \big(\int_{\overline{h}_0(\overline{t}, \overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \overline{\boldsymbol{q}}_0 \big) + \boldsymbol{q}_0 \cdot \nabla_{\overline{x}} \overline{h}_{\rm soil}) |_{\overline{\Gamma}_{\rm soil}} = 0$ on $\overline{\Gamma}_{\rm soil}$. It follows that $\overline{\boldsymbol{q}}_1 \cdot \boldsymbol{e}_3 = 0$ in $]0, \overline{T}[\times \overline{\Omega}$ (see (A.86)). The second equation of (A.80) gives

$$\overline{\boldsymbol{q}}_1 = -\frac{\kappa(\overline{P}_0)}{\rho\, g} \boldsymbol{G}_0 \nabla_{\overline{x}} \overline{P}_0 = -\kappa(\overline{P}_0) \boldsymbol{G}_0 \nabla \overline{H}_0 \qquad \text{in }]0, \overline{T}[\times \overline{\Omega}.$$

The characterization of $\overline{H}_0$ depends on the values of α. Similar arguments to those employed in the intermediate time case when $\alpha \neq 0$ lead to (A.59).

It remains to consider the case $\alpha = 0$. In this case, we first remark that the compatibility condition $\overline{F}_0 = 0$ is imposed (see (A.85) for $k = 0$). Thus, using (A.81) and (A.82), we obtain

$$\overline{\boldsymbol{v}}_1 = \overline{\boldsymbol{q}}_1 + \overline{\boldsymbol{w}}_1 = -\kappa(\overline{P}_0) \big(\overline{\boldsymbol{A}}_0 + \overline{\boldsymbol{G}}_0 \big) \nabla \overline{H}_0.$$

In particular, (A.62) thanks to $\boldsymbol{M}_0 = \boldsymbol{A}_0 + \boldsymbol{G}_0$. Moreover the first equation in (A.85) for $k = 1$ implies $\overline{F}_1 = 0$ (as $\alpha = 0$). It remains to obtain the first relation of the system (A.61). By substituting the asymptotics (A.53)–(A.56) into the second equation in (A.90) and by identifying the coefficients associated with ε^2, we obtain (*cf.* (A.80))

$$\begin{aligned} -\nabla' \cdot \big(\overline{\boldsymbol{J}}(\overline{H}_0)\, \nabla' \overline{H}_0 \big) = -(\overline{\boldsymbol{q}}_2 \cdot \boldsymbol{e}_3)|_{\overline{\Gamma}_{\rm soil}} + (\overline{\boldsymbol{q}}_1 \cdot \nabla \overline{h}_{\rm soil})|_{\overline{\Gamma}_{\rm soil}} - \Big(\int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \partial_{\overline{t}} \theta(\overline{P}_0) + \alpha_p \theta(\overline{P}_0)\, \partial_{\overline{t}} \overline{P}_0 \, d\overline{z} \Big) \\ - \nabla_{\overline{x}} \cdot \Big(\int_{\overline{h}_0(\overline{t}, \overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \overline{\boldsymbol{q}}_1 \Big) \quad \text{for } (\overline{t}, \overline{\boldsymbol{x}}) \in]0, \overline{T}[\times \overline{\Omega}_x. \end{aligned}$$

We note that by the equality $\alpha = 0$ and by the first equation in (A.85) for $k = 2$, we have $(\overline{\boldsymbol{q}}_2 \cdot \boldsymbol{e}_3)|_{\overline{\Gamma}_{\rm soil}} - (\overline{\boldsymbol{q}}_1 \cdot \nabla \overline{h}_{\rm soil})|_{\overline{\Gamma}_{\rm soil}} = \overline{F}_2 / \beta$. It comes

$$-\nabla' \cdot \left(\Big(\overline{\boldsymbol{J}}(\overline{H}_0) + \int_{\overline{h}_0}^{\overline{h}_{\rm soil}} \kappa(\overline{P}_0) \boldsymbol{N}_0 \Big) \nabla' \overline{H}_0 \right) = -\frac{\overline{F}_2}{\beta} - \Big(\int_{\overline{h}_{\rm bot}(\overline{\boldsymbol{x}})}^{\overline{h}_{\rm soil}(\overline{\boldsymbol{x}})} \partial_{\overline{t}} \theta(\overline{P}_0) + \alpha_p \theta(\overline{P}_0)\, \partial_{\overline{t}} \overline{P}_0 \, d\overline{z} \Big)$$

$$\text{for } (\overline{t}, \overline{\boldsymbol{x}}) \in]0, \overline{T}[\times \overline{\Omega}_x.$$

Hence, the latter relation combined with the third boundary condition in (A.85) and the definition (A.44) justify the problem (A.61). □

[a] Univ. du Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France. [b] CNRS FR 2037, France. [c] Université Libanaise, LaMA-Liban, Laboratoire de recherche en Mathématiques et Applications, P.O. Box 37 Tripoli, Liban.
Email address: Safaa.Al-Nazer@etu.univ-littoral.fr

[1] Univ. Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France. [b] CNRS FR 2037, France
Email address: Christophe.Bourel@univ-littoral.fr

[a] Univ. du Littoral Côte d'Opale, UR 2597, LMPA, Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville, F-62100 Calais, France.[b] CNRS FR 2037, France.
Email address: Carole.Rosier@univ-littoral.fr

Chapitre 5

Conclusion

Dans la première partie du travail, on a présenté plusieurs solveurs chimiques stables et précis à intégrer dans un algorithme séquentiel itératif pour le transport réactif. Les méthodes présentées permettent de résoudre les équilibres thermodynamiques d'une manière totalement nouvelle (sans utiliser la méthode de Newton-Raphson). Ces solveurs chimiques combinent la méthode des fractions continues positives (FCP) à deux méthodes numériques itératives, la méthode d'Accélération d'Anderson (AA) et des méthodes d'extrapolation vectorielle, à savoir les méthodes MPE (minimal polynomial extrapolation) et RRE (reduced rank extrapolation). Le principal avantage de ces approches est d'éviter de former la matrice jacobienne et donc d'éviter les problèmes liés aux mauvais conditionnements de la matrice qui sont observés lors des résolutions classiques résultant de la méthode de Newton-Raphson.
Une étude expérimentale met en œuvre les trois méthodes numériques pour le problème d'équilibre thermodynamique notamment sur le cas test de l'acide gallique et pour le cas test 1D "Easy" du Benchmark MoMas. Les résultats numériques présentés améliorent les résultats existants (par exemple ceux donnés dans [22]). Concernant le cas test de l'acide gallique, les résultats sont comparés aux résultats de Jérôme Carrayrou issus des méthodes de type Newton-Raphson, développés dans [29]. Pour le cas test 1D "Easy" de Benchmark MoMas, on compare nos résultats aux résultats obtenus par certains codes de transport réactifs qui ont participé à la réalisation du Benchmark. Parmi ces codes, on cite le code HYTEC [19], où toutes les réactions chimiques sont résolues par le code de spéciation CHESS [18] qui utilise un schéma amélioré de Newton-Raphson pour résoudre l'ensemble des équations algébriques non linéaires décrivant le système chimique. On donne également une comparaison de nos résultats avec ceux obtenus par d'autres codes [39] comme SPECY, MIN3P, GDAE et Hoffmann et al, qui sont basés sur une méthode de type Newton pour linéariser le système chimique et qui utilisent chacun une méthode spécifique pour trouver la solution du système d'équations linéarisé. Ces comparaisons assurent la rapidité, la robustesse ainsi que la stabilité des méthodes (AA), (MPE) et (RRE) et montrent aussi et surtout qu'elles ont une meilleure performance en terme de temps CPU.

Evidemment, la parallélisation potentielle des algorithmes proposés dans cette partie est une étape importante dans les travaux à venir, en particulier si ces algorithmes devaient être implémentés dans le cadre d'une plateforme open source parallèle. La parallélisation des algorithmes MPE et RRE a déjà été discutée, notamment dans le cadre de l'article [111]. Il semble tout à fait possible d'adapter ces résultats à notre cas.

Dans la seconde partie de la thèse, on s'est intéressé à l'aspect "écoulement engendrant le transport" de la problématique du transport réactif en milieu poreux.
Notre étude est partie du modèle établi dans [48], qui est une alternative au problème de Richards $3D$ pour décrire l'écoulement dans un aquifère peu profond dans une large gamme d'échelles de temps. La première remarque concernant ce modèle est qu'il a deux avantages par rapport au modèle de

Richards $3D$ dont il est issu :

— Premièrement il engendre des problèmes numériques plus rapides à résoudre puisqu'il résulte du couplage d'un problème $2D$ avec une multitude de problèmes de Richards $1D$ verticaux qui pourront être résolus en parallèle offrant ainsi un gain de temps important dans le calcul numérique.

— Deuxièment, ce problème couplé et le problème original de Richards $3D$ présentent les mêmes comportements dominants lorsque le *rapport* ϵ = profondeur/largeur de l'aquifère est petit. De plus ces comportements effectifs sont identiques pour tous les choix d'échelles de temps considérés.

Mais l'analyse mathématique de ce modèle s'avère très délicate et on peut en énumérer les principales difficultés comme suit :

— l'intégration des équations sur un domaine à frontière libre,

— la nonlinéarité et la dégénérescence apparaissant dans les dérivées en temps des deux équations,

— la perte de contrôle des composantes horizontales de la pression.

Nous introduisons deux modèles $(\mathcal{M})$ et $(\mathcal{N})$ pour lesquels on pourra établir des résultats d'existence dans des espaces convenables et dont les résultats numériques sont espérés être proches de ceux présentés dans [48].

Usuellement, la transformée de Kirchoff est appliquée à l'équation de Richards (sous des hypothèses appropriées sur la porosité et la perméabilité) pour éliminer la nonlinéarité dans le terme diffusif. Dans ce travail, nous préférons exploiter l'hypothèse de la faible compressibilité de l'eau qui permet d'éliminer les dégénérescences et nonlinéarités présentes dans les dérivées temporelles des nouveaux systèmes. Ces transformations nous ramènent au cadre des équations paraboliques quasi-linéaires sur des domaines non cylindriques auxquelles nous pouvons appliquer la méthode des domaines auxiliaires introduite par Lions et Mignot [57, 58] pour traiter les deux problèmes à frontière libre $(\mathcal{M})$ et $(\mathcal{N})$ et ainsi résoudre les deux premières difficultés précédemment citées.
Le manque de contrôle du gradient de la pression par rapport aux variables horizontales du modèle établi dans [48] résulte essentiellement d'une conductivité hydraulique horizontale prise nulle dans la partie insaturée de l'aquifère (ce qui permet de résoudre plusieurs verticaux $1D$). Nous reintroduisons une faible conductivité hydraulique horizontale dans cette zone de l'aquifère, ce qui nous permet de pallier à cette perte d'information sur les composantes horizontales du gradient de pression.

Mentionnons maintenant les différences entre les deux modèles :

Pour l'étude du modèle $(\mathcal{M})$, nous nous plaçons dans le cas isotrope ce qui nous permet d'introduire une conductivité hydraulique moyennée verticalement. La moyenne verticale des lois de conservation dans la zone de saturation conduit alors à une équation elliptique dégénérée dont la dégénérescence dépend de l'épaisseur de la zone saturée. Par ailleurs, la prise en compte de la compressibilité de l'eau introduit une dégénérescence dans la dérivée temporelle qui dépend aussi de l'épaisseur de la zone saturée. Un changement de variable permet alors d'absorber les deux termes dégénérés et de revenir à une équation parabolique régulière.
Pour le modèle $(\mathcal{N})$, nous considérons une forme très générale pour la conductivité hydraulique (comme cela a été fait dans [48]), ce qui induit une nonlinéarité supplémentaire dans la dérivée temporelle de l'équation régissant l'évolution de l'interface (équation en h). Cela change donc totalement son étude puisqu'on ne peut plus faire le changement de variable qui avait permis pour le modèle $(\mathcal{M})$ de neutraliser les deux nonlinéarités. De plus, contrairement au modèle $(\mathcal{M})$, nous considérons une condition

aux limites de Robin à l'interface Γ_{soil} permettant ainsi de décrire les échanges entre les eaux de surface et les eaux souterraines mais induisant donc des termes supplémentaires dans le système final. Enfin, nous construisons le modèle $(\mathcal{N})$ de sorte à préserver la conservation de la masse ainsi que le faisaient les modèles décrits dans [48].

Concernant les perpectives, on pourrait les classer en deux catégories selon qu'elles sont à court terme ou long terme. Commençons par les perpectives à court terme :

— Pour le problème des équilibres thermodynamiques, il serait vraiment très intéressant de comparer les résultats obtenus par les approches AA, RRE ou MPE avec ceux obtenus grâce aux méthodes *deep learning* utilisées dans [104, 105].

— Pour finaliser l'étude mathématique des modèles $(\mathcal{M})$ et $(\mathcal{N})$, il faudrait ajouter les simulations numériques associées aux deux modèles de sorte de les comparer à celles obtenues dans [48]. Soulignons que ces deux études ont fait apparaître une nouvelle inconnue u (pour laquelle la positivité est démontrée) qui induit une construction de l'inconnue h directement dans l'intervalle d'étude souhaité (h_{bot}, h_{soil}), ce qui n'était pas automatiquement garanti pour les solutions des modèles décrits dans [48].

— Nous avons considéré dans chacun des modèles $(\mathcal{M})$ et $(\mathcal{N})$ une faible conductivité hydraulique horizontale dans la partie insaturée de l'aquifère, ce serait intéressant de voir si nous pouvons montrer que les solutions limites obtenues lorsque cette conductivité hydraulique horizontale tend vers zéro, coincident avec celles des modèles décrits dans [48] au moins numériquement dans un premier temps, puis théoriquement dans un second temps.

Mais le vrai challenge qui constitue le projet à long terme est de coupler les deux parties de ce travail, la chimie et le transport afin de proposer de nouveaux solveurs pour la résolution numérique du transport réactif en milieu poreux, qui seraient numériquement stables et de par leur conception permettraient de réaliser des gains de temps CPU notables.

Bibliographie

[1] Anderson, D. G. *Iterative procedures for nonlinear integral equations.* Journal of the ACM, 12 (1965), pp. 547-560.

[2] Sidi, A. *Convergence and Stability Properties of Minimal Polynomial and Reduced Rank Extrapolation Algorithms.* SIAM J. Numer. Anal. 23, no.1, pp. 197-209, 1986.

[3] Sidi, A. *Vector Extrapolation methods with applications to solution of large systems of equations and to PageRank computations.* Computer science department, Technion - Israel institute of technology, Haifa 32000, Israel.

[4] Sidi, A. *Extrapolation vs. projection methods for linear systems of equations.* J. Comput. Appl. Math. 22, pp. 71-88, 1988.

[5] Sidi, A. *Efficient implementation of minimal polynomial and reduced rank extrapolation methods.* J. Comput. Appl.Math. vol. 36 (1991), p. 305-337 (cf. p. 32 ?34).

[6] Jbilou, K. and Sadok, H. *Vector extrapolation methods. Application and numerical comparison* J. Comp. Appl. Math, 122(2000), pp. 149-165.

[7] Jbilou, K. and Sadok, H. *Analysis of some vector extrapolation methods for linear systems.* Numerische Mathematic, 70(1995), pp. 73-89.

[8] Jbilou, K. and Sadok, H. *Some results about vector extrapolation methods and related fixed point iteration.* J. Comp. Appl. Math., 36(1991), pp. 385-398.

[9] Jbilou, K. *A general projection algorithm for solving linear systems of equations.* Numer. Algorithms, 4 (1993), pp. 361-377.

[10] Davis, T. A. and Hu, Y. *The University of Florida SparseMatrix Collection.* ACM T.Math. Software vol. 38 (2011), pp. 1-25 (cf. p. 147).

[11] Sadok, H. *About Henrici's transformation for accelerating vector sequences.* J. Comput. Appl.Math. 29 (1990) 101-110.

[12] Cabay, S. and Jackson, L.M. *A polynomial extrapolation method for finding limits and antilimits of vector sequences.* SIAM J. Numer. Anal. 13 (1976) 734-752.

[13] Eddy, R.P. *Extrapolating to the limit of a vector sequence.* P.C.C. Wang, Ed., Information Linkage between Applied Mathematics and Industry (Academic Press, New York, 1979) 387-396.

[14] MeSina, M. *Convergence acceleration for the iterative solution of the equations* $X = AX + f$,. Comput. Methocis Appl. Mech. Engrg. 10 (2) (1977) 165-173.

[15] Parkhurst, D. L. and Appelo, C. A. J. *A computer program for speciation, batch-reaction, one-dimensionnel transport, and inverse geochemical calculations.* Technical Report 99-4259 USGS, (1990).

[16] Parkhurst, D. L. and 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. Rep. 99-4259, U.S. Geological Survey, Denver, CO (1999).

[17] Van Der Lee, J. *Modélisation du comportement géochimique et du transport des radionucléides en présence de colloides.* Thèse de l'Ecole Nationale Supérieure des Mines de Paris, (1997).

[18] Van Der Lee, J. *Thermodynamic and mathematical concepts of CHESS*. Technical Report LHM/RD/98/39, CIG- Ecole des l'vlines de Paris, Fontainebleau, France, (1998).

[19] Lagneau, V. and Van Der Lee, J. *HYTEC results of the MoMas reactive transport benchmark*.

[20] Westall, C., Zachary, J. L. and Morel, F. M. M. *A computer program for the calculation of chemical equilibrium composition of aqueous system*. Massachusetts Institute of Technology Technical Note, 18 (1976).

[21] Rubini, J. *Transport of reacting solutes in porous media : Relation between mathematical nature of problem formulation and chemical nature of reactions.* Water Resources Research, 19 :1231,1252, (1983).

[22] Carrayrou, J., Mosé, R. and Behra, P. *New efficient algorithm for solving thermodynamic chemistry.* AIChE J. 48(4), 894-904 (2002).

[23] Stumm, W. and Morgan, J. J. *Aquatic chemistry. Chemical equilibria and rates in natural waters.* Third edition, Wiley-interscience New York, pp 1022. (1996).

[24] Steefel, C. I. and 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).

[25] Steefel, C. I., Appelo, C. A. J., Arora, B., Jacques, D., Kalbacher, T., Kolditz, O., Lagneau, V., Lichtner, P. C.,Mayer, K.,Meeussen, J. C. L., Molins, S., Moulton, D., Shao, H., Simunek, J., Spycher, N., Yabusaki, S. B. and Yeh, G. T. *Reactive transport codes for subsurface environmental simulation*. Comput Geosci. (2015) ; 19(3) : 445-478.

[26] Bethke, C. M. *Geochemical and Biogeochemical Reaction Modeling*. Cambridge University Press, 2nd edition edition, (2008).

[27] Morel, F.M.M. and Hering, J.G. *Principles and Applications of Aquatic Chemistry.* Wiley, New-York, 2nd edition, (1993).

[28] Shapiro, N. Z. and Shapley, L. S. *Mass action laws and the Gibbs free energy function*. J. SOC. Indust. Appl. Math., 13(2) :353-375, (1965).

[29] Carrayrou, J. *Modélisation du transport de solutés réactifs en milieu poreux saturé.* Thèse de doctorat, Université Louis Pasteur, Strasbourg, (2001).

[30] Krebs, R., Sardin, M. and Schweich, D. *Mineral Dissolution, Precipitation and Ion Exchange in Surfactant Flooding*. AIChE J., 33, 1371 (1987).

[31] Wigley, T. M. L. *WATSPEC : A Computer Program for Determining the Equilibrium Speciation of Aqueous Solutions*. Brit. Geomorphol. Res. Group Tech. Bull. 20 (1977).

[32] Walker, H.F. and Ni, P. *Anderson acceleration for fixed-point iterations*. SIAM J. Numer. Anal. 49 (4) (2011) 1715-1735.

[33] Walker, H.F. *Anderson Acceleration : Algorithms and Implementations*. Research Report, MS-6-15-50, Worcester Polytechnic Institute Mathematical Sciences Department, (2011).

[34] Toth, A. and Kelley, C. T. *Convergence analysis for Anderson acceleration*. SIAM J. Numer.Anal., 53 (2015), pp. 805-819. https://doi.org/10.1137/130919398.

[35] Fang, H.and Saad, Y. *Two classes of multisecant methods for nonlinear acceleration*. Numer. Linear Algebra Appl. 16 (3) (2009) 197-221.

[36] Brassard, P. and Bodurtha, P. *A Feasible Set for Chemical Speciation Problems*. Comput. Geosci., 26, 277 (2000).

[37] Marinoni M., Carrayrou, J., Lucas Y., Ackerer P., *Thermodynamic equilibrium solutions through a modified Newton Rphson method*, AIChE Journal (2016).

[38] Machat, H. and Carrayrou, J. *Comparison of linear solvers for equilibrium geochemistry computations*. Computational Geosciences, (2017). 21(1) : p. 131-150.

[39] Carrayrou, J., Hoffman, J., Knabner, P., Krautle,S. , De Dieuleveult, C., Erhel, J., Van Der Lee, J., Lagneau, V., Mayer, K. U. and Macquarrie, K. T. B. *Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems-the MoMaS benchmark case*. Computat. Geosci., 14 (2010), pp. 483-502.

[40] Carrayrou, J., Kern, M. and Knabner, P. *Reactive transport benchmark of MoMaS*. Computation Geosci., 14 (2010), pp. 385-392. 10.1007/s10596-009-9157-7.

[41] Carrayrou, J. *Looking for some reference solutions for the reactive transport benchmark of MoMaS with SPECY*. Computational Geosciences, 14 :393-403 (2010).

[42] Jacob, B. *Dynamics of fluids in porous media*. Elsevier, New-York (1972).

[43] Jacob, B.and Arnold, V. *Modeling groundwater flow and pollution*. Springer, Netherlands (1987).

[44] Richardson, L. F. *Weather prediction by numerical process*. Cambridge, The University press, page 262, (1922).

[45] Ackerer, P. and Younes, A. *Efficient approximations for the the simulation of density driven flow in porous media*. Adv. Water Res., Vol. 31, 15-27, (2008).

[46] Genuchten, M. T. V. *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils 1*. Soil science society of America journal, 44(5) : 892-898, 1980.

[47] Brooks, R. H. and Corey, A. T. *Hydraulic properties of porous media*. Colorado State University, 1964. Hydrology Papers. Soil Science society of Americal journal, 44(5) : 892-898, 1980.

[48] Bourel, C., Choquet, Rosier, C. and Tsegmid, M. *Modelling of shallow aquifers in interaction with overland water.* Submitted.

[49] Tsegmid, M. *Modélisation d'aquifères peu profonds en interaction avec les eaux de surfaces*. Phd Thesis, 2019 ULCO. www.theses.fr/2019DUNK0526

[50] Alt, H. W. and Luckhaus, S. *Quasilinear elliptic-parabolic differential equations*. Math. Z., Vol. 1, 311–341, 1983.

[51] Alt, H. W. and Di Benedetto, E. *Nonsteady flow of water and oil through inhomogeneous porous media, Ann. Scuola Norm. Sup. Pisa 12 (1985) 335-392.*

[52] Chen, X., Friedman, A. and Kimura, T. *Nonstationary filtration in partially saturated porous medium*. Euro. J. Appl. Math. 5 (1994) 405-429.

[53] Hulshof, J. and Wolanski, N. *Monotone flows in N-dimensional partially saturated porous media : Lipschitz continuity of the interface*. Arch. Rat. Mech. Anal. 102 (1988) 287-305.

[54] Hoffmann J., Kräutle S. and Knaber P., *A parallel global-implicit @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 (2010) 421-433.

[55] Showalter, R. E. and Su, N. *Partially saturated flow in a poroelastic medium*. Disc. Cont. Dyn. Syst. Ser. B (2001) 403-420.

[56] Yin, H. M. *A singular-degenerate free boundary problem arising from the moisture evaporation in a partially saturated porous medium*. Ann. Mat. Pura Appl. **161** (1992) 379-397.

[57] Lions, J. L. *Sur les problèmes mixtes pour certains systèmes paraboliques dans des ouverts non cylindriques*. Annales de l'Institut Fourier 7, 1957, p. 143-182.

[58] Mignot, A. L. *Méthodes d'approximation des solutions de certains problèmes aux limites linéaires.* Rendiconti del Seminario Matematico della Università di Padova, tome 40 (1968), p. 1-138.

[59] Ladyzenskaja, O. A., Solonnikov, V.A. and Ural'ceva, N. N. *Linear and quasilinear equations of parabolic type.* Translated from the Russian by S. Smith. Translations of Mathematical Monographs, Vol. 23. American Mathematical Society, Providence, R.I., 1968.

[60] Lions, J. L., and Magenes, E. *Problèmes aux limites non homogènes.* Vol. 1, Dunod, 1968.

[61] Simon, J. *Compact sets in the space $L^p(0,T,B)$.* Ann. Mat. Pura Appl., vol. 146 (4), 65–96, 1987.

[62] Gagneux, G. and Madaune-Tort, M. *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière.* Mathématiques & Applications, 22, Springer, 1996.

[63] Evans, L. C. *Partial differential equations.* Volume 19 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, second edition, 2010.

[64] Zeidler, E. *Nonlinear functional analysis and its applications.* Part 1, Springer Verlag, 1986.

[65] Pantelis, G. *Saturated-unsaturated flow in unconfined aquifers.* Zeitschrift für angewandte Mathematik und Physik ZAMP, 36(5) :648-657, Sep 1985.

[66] Jazar, M. and Monneau, R. *Derivation of seawater intrusion models by formal asymptotics.* SIAM J. Appl. Math., 74(4) :1152-1173, 2014.

[67] Al Nazer, S., Rosier, C. and Tsegmid, M. *Derivation and mathematical analysis of dupuit-richards model taking into account the fluid compressibility.* submitted, 2020.

[68] Al Nazer, S., Bourel, C., and Rosier, C. *Global existence result for a coupled system describing the exchanges between a shallow aquifer and the overland water.* submitted, 2020.

[69] Richards, L. A. *Capillary conduction of liquids through porous mediums.* Physics 1(5) : 318-333, 1931.

[70] Bourgeat, A., Bryant, S., Carrayrou, J., Dimier, A., Van Duijn, C.J., Kern, M. and Knabner, P. *Benchmark Reactive Transport.* Technical Report GDRMOMAS (2006).

[71] Brezinski, C., Redivo Zaglia, M. and Saad, Y. *Shanks sequence transformations and Anderson acceleration.* SIAM Rev., 60 (2018) 646-669.

[72] Broyden, C. G. *A class of methods for solving nonlinear simultaneous equations.* Math. Comp. 19, 577-593 (1965).

[73] Eyert, V. *A comparative study on methods for convergence acceleration of iterative vector sequences.* J. Comput. Phys. 124(2), 271-285 (1996).

[74] Morin, K. A. *Simplified Explanations and Examples of Computerized Methods for Calculating Chemical Equilibrium in Water.* Comput. Geosci., 11, 409 (1985).

[75] Nelder, J. A. and Mead, R. *A Simplex method for functionMinimization.* Comput. J., 7, 308 (1965).

[76] Potra, F. A. and Engler, H. *A characterization of the behavior of the Anderson acceleration on linear problems.* Linear Algebra Appl. 438(3), 1002-1011 (2013).

[77] Pulay, P. *Convergence acceleration of iterative sequences. The case of SCF iteration.* Chem. Phys. Lett. 73(2), 393-398 (1980).

[78] Rohwedder, T. and Schneider, R. *An analysis for the DIIS acceleration method used in quantum chemistry calculations.* J.Math. Chem. 49(9), 1889-1914 (2011).

[79] Saad, Y. and Schultz, M.H. *GMRES : a generalized minimal residual algorithm for solving nonsymmetric linear systems.* SIAM J. Sci. Statist. Comput. 7(3), 856-869 (1986).

[80] Saaltink, MW., Carrera, J. and Ayora, C. *Comparison of two approaches for reactive transport modeling.* J. Geochem Explor. (2000) ; 69 :97- 101.

[81] Wood, J. R. *Calculation of Fluid-Mineral Equilibria Using the Simplex Algorithm.* Comput. Geosci., 19, 23 (1993).

[82] Yeh, G. T. and Tripathi, V. S. *A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components.* Water Resources Res. 25, 93-108, (1989).

[83] Abbott, M. B., Bathurst, J. C., Cunge, JA. O'connell, P. E., and Rasmussen, J. *An introduction to the european hydrological system - système hydrologique europeen, "she", 2 : Structure of a physically-based, distributed modelling system.* Journal of Hydrology, 87(1) : 61-77, 1986.

[84] Alkhayal, J., Issa, S., Jazar, M. and Monneau, R. *Existence results for degenerate cross-diffusion systems with application to seawater intrusion.* ESAIM Control Optim. Calc. Var. 24, no. 4,1735-1758 (2018).

[85] Bensoussan, A., Lion, J. L. and Papanicoulou, G. *Asymptotic analysis for periodic structure, North-Holland, Amsterdam, 1978.*

[86] Bernardi, C., Blouza, A. and El Alaoui, L. *The rain on underground porous media part i : Analysis of a richards model.* Chinese Annals of Mathematics, Series B, 34(2) : 193-212,Mar 2013.

[87] Berninger, H., Ohlberger, M., Sander, O. and Smetana, K. *Unsaturated subsurface flow with surface water and nonlinear in- and outflow conditions.* Mathematical Models and Methods in Applied Sciences, 24(05) : 901-936, 2014.

[88] Benilan, P., Boccardo,L., Gallouet, T., Gariepy, R., Pierre, M. and Vazquez, J.L. *An L^1 theory of existence and uniqueness of nonlinear elliptic equations.* Ann. Sc. Norm. Super. Pisa, Cl. Sci., IV. Ser., 22 (1995), 240-273.

[89] Choquet, C., Diédhiou, M. M.and Rosier, C. *Derivation of a Sharp-Diffuse Interfaces Model for Seawater Intrusion in a Free Aquifer.* Numerical Simulations, SIAM J. Appl.Math. 76 (2016), no. 1, 138-158.

[90] Darcy, H. *Les fontaines publiques de la ville de Dijon ; exposition et application des principes à employer dans les questions de distribution d'eau.* Victor Dalmont, Editeur, Paris, (1856).

[91] Dupuit, J. 1863. *Etudes théoriques et pratiques sur le mouvement des eaux dans les canaux découverts et à travers les terrains perméables.* 2ème édition, Dunod, Paris, (1863).

[92] Sochala, P., Ern, A. and Piperno, S. *Mass conservative bdf-discontinuous galerkin/explicit finite volume schemes for coupling subsurface and overland flows.* Computer Methods in AppliedMechanics and Engineering, 198(27) :2122 - 2136, 2009.

[93] Fetter, C.W. *Hydrogeology : A shot history, Part 2.* GroundWater, 42 (2004), 949-953.

[94] Hulshof, J. and Wolanski, N. *Monotone flows in N-dimensional partially saturated porous media : Lipschitz continuity of the interface.* Arch. Rat.Mech. Anal. 102 (1988) 287-305. Mathématiques and Applications, 22, Springer, 1996.

[95] Kong, J., Xin, P., Song, Z. and Li, L. *A new model for coupling surface and subsurface water flows :With an application to a lagoon.* Journal of Hydrology, 390(1) : 116 - 120, 2010.

[96] Marle, M. C. *Henry Darcy et les écoulements de fluides en milieux poreux, Oil and Gas science and technology* Re. IFP, Vol. 61 (5) (2006), 599-609.

[97] Meyers, N. G. *An L^p-estimate for the gradient of solution of second order elliptic divergence equations.* Ann. Sc. Norm. Sup. Pisa, Vol. 17, pp. 189-206, 1963.

[98] Q Pham, H., GFredlund, D. and Lee Barbour, S. *A study of hysteresis models for soil-water characteristic curves.* Canadian Geotechnical Journal, 42(6) : 1548-1568, 2005.

[99] F Pikul, M., L Street, R. and Remson, I. *A numerical model based on coupled one-dimensional richards and boussinesq equations.* Water Resources Research, 10(2) : 295-302, 1974.

[100] Rosier, C. and Rosier, L. *Well-posedness of a degenerate parabolic equation issuing from two-dimensional perfect fluid dynamics.* Applicable Anal., Vol. 75 (3-4), pp 441-465, 2000.

[101] Schweizer, B. *Hysteresis in porous media : Modelling and analysis.* Interfaces and Free Boundaries. 19 : 417-447, 01 2017.

[102] Bénilan, P., Boccardo, L., Gallouet, T., Gariepy, R., Pierre, M. and Vazquez, L. L. *An L^1-theory of existence and uniqueness of solutions of nonlinear elliptic equations.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 22(2) : 241-273, 1995.

[103] Zhang, T., Li, Y. and Sun, S. *Phase equilibrium calculations in shale gas reservoirs. Capillarity*, 2019, 2(1) : 8-16, doi : 10.26804/capi.2019.01.02.

[104] Zhang, T., Li, Yu, Li Yiteng, Sun, S. and Hua B. *A self-adaptive deep learning algorithm for accelerating multicomponent flash calculation*. Computer Methods in Applied Mechanics and Engineering (IF 5.763) Volume 369 (2020), DOI : 10.1016/j.cma.2020.113207.

[105] Zhang, T., Li Yiteng, Sun, S.and Gao X. *Accelerating flash calculations in unconventional reservoirs considering capillary pressure using an optimized deep learning algorithm*. Journal of Petroleum Science and Engineering, Vol 195 (2020).

[106] Ackerer, P., Preface. *Special issue on simulations of reactive transport : Results of the MoMaS benchmarks*, Computat. Geosci. 14(3) (2010), 383.

[107] Mayer, K.U. and MacQuarrie, K.T.B. *Solution of the MoMaS reactive transport benchmark with MIN3P-model formulation and simulation results*, Comput. Geosci. 14 (2010) 405-419.

[108] Ahusborde, E., Ossmani, M. E. and Id Moulay, M., *A fully implicit finite volume scheme for single phase flow with reactive transport in porous media*, Mathematics and computers in simulation 164 (2019) 3-23.

[109] DuMuX, DUNE for Multi-Phase, Component, Scale, Physics, ..., flow and transport in porous media.

[110] DUNE. the Distributed and Unified Numerics Environment. http://www.dune.project.org.

[111] Duminil, S., Sadok, H. and Silvester, D. *Fast solvers of discretized Navier-Stokes problems using vector extrapolation*. Numer. Algorithmes 66 (2014), no. 1, 89-104.

Annexe A

Codes de programmation avec le logiciel Matlab R2018a

Nous présentons dans cette partie les codes numériques utilisés pour chercher la solution du système algébrique non linéaire qui décrit l'équilibre thermodynamique en milieu poreux.

Définition du cas test d'acide gallique

```
function [Asolution,Ksolution,T,sol]=acide_guallic_test

T=[1.e-3;1.e-3];pH=5.8;
%initial guess:i1=log10([10.^-11;5*10.^-4]);i2=log10([5.012*10.^-10;10.^-9]);
[KSOLUTION,~,ASOLUTION,~,~]=get_equilib_defn;
for i=1:size(pH,2)
    %adjust for fixed pH
    [Ksolution,Asolution]=get_equilib_pH(KSOLUTION,ASOLUTION,pH(i));
    %number of components
    Nc=size(Asolution,2);
    %number of secondary species
    Ne=size(Asolution,1) ;    %#ok<*NASGU>
End
sol=[-4.6930;-6.5870];

end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 function [KSOLUTION,lnK,ASOLUTION,SOLUTIONNAMES,C]=get_equilib_defn

KSOLUTION=[0;0;0;-14;-4.15;-12.59;-23.67;-4.93;-9.43;-21.98;-37.69;-
22.65;-27.81;-32.87;-39.56;-20.25;-12.52] ;
K=10.^KSOLUTION;
lnK=log(K);
ASOLUTION=[1 0 0;0 1 0;0 0 1;-1 0 0;-1 0 1;-2 0 1;-3 0 1;-2 1 1;-3 1
1;-6 1 2;-9 1 3;-8 2 3;-9 2 3;-10 2 3;

    -11 2 3;-9 4 3;-5 3 1];
 C=[1.58*10.^-6;2.03*10.^-5;2.59*10.^-7;6.31*10.^-9;1.16*10.^-
5;2.65*10.^-8;1.39*10.^-13;2.45*10.^-5;
    4.90*10.^-4;8.97*10.^-6;1.14*10.^-10;4.01*10.^-6;1.75*10.^-
5;9.61*10.^-5;1.24*10.^-4;2.61*10.^-7;6.51*10.^-5];
SOLUTIONNAMES=strvcat('H','Al','H3L','OH','H2L','HL','L','AlHL','AlL',
'AlL2','AlL3','S1','S2','S3','S4','S5','S6'); %#ok<DSTRVCT>

end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [Ksolution,Asolution]=get_equilib_pH(KSOLUTION,ASOLUTION,pH)

[~,Nc]=size(ASOLUTION);
Ksolution=KSOLUTION-ASOLUTION(:,1)*pH;
Asolution=ASOLUTION(:,2:Nc);

end
```

Définition du cas test « 1D easy » de Benchmark MoMas

```
% dans la zone A
function [Asolution,Ksolution,TA,solA]=momas_easy_A_test

Asolution=[1 0 0 0 0 ;0 1 0 0 0 ;0 0 1 0 0 ;0 0 0 1 0 ;0 -1 0 0 0 ;0 1
1 0 0 ;0 -1 0 1 0 ;0 -4 1 3 0 ;
    0 4 3 1 0;0 0 0 0 1;0 3 1 0 1;0 -3 0 1 2 ];
Ksolution=[0;0;0;0;-12;0;0;-1;35;0;6;-1];
TA=[10.^-20;-2;10.^-20;2;1];
solA=log10([10.^-20;0.25972;10.^-20;0.34954;0.39074]);

end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% dans la zone B
function [Asolution,Ksolution,TB,solB]=momas_easy_B_test

Asolution=[1 0 0 0 0 ;0 1 0 0 0 ;0 0 1 0 0 ;0 0 0 1 0 ;0 -1 0 0 0 ;0 1
1 0 0 ;0 -1 0 1 0 ;0 -4 1 3 0 ;
    0 4 3 1 0;0 0 0 0 1;0 3 1 0 1;0 -3 0 1 2 ];
Ksolution=[0;0;0;0;-12;0;0;-1;35;0;6;-1];
TB=[10.^-20;-2;10.^-20;2;10];
solB=log10([10.^-20;1.5116;10.^-20;0.57561;7.9128]);

end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Periode d'injection %t<5000s
function [Asolution,Ksolution,TI,solI]=momas_easy_inj_test

Asolution=[1 0 0 0 0 ;0 1 0 0 0 ;0 0 1 0 0 ;0 0 0 1 0 ;0 -1 0 0 0 ;0 1
1 0 0 ;0 -1 0 1 0 ;0 -4 1 3 0 ;
    0 4 3 1 0;0 0 0 0 1;0 3 1 0 1;0 -3 0 1 2 ];
Ksolution=[0;0;0;0;-12;0;0;-1;35;0;6;-1];
solI=log10([0.3;0.24162;0.24162;10.^-50;10.^-23]);
TI=[0.3;0.3;0.3;10.^-20;10.^-20];

end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Periode de lessivage %t>5000s
function [Asolution,Ksolution,TL,solL]=momas_easy_leach_test

Asolution=[1 0 0 0 0 ;0 1 0 0 0 ;0 0 1 0 0 ;0 0 0 1 0 ;0 -1 0 0 0 ;0 1
1 0 0 ;0 -1 0 1 0 ;0 -4 1 3 0 ;
    0 4 3 1 0;0 0 0 0 1;0 3 1 0 1;0 -3 0 1 2 ];
Ksolution=[0;0;0;0;-12;0;0;-1;35;0;6;-1];
TL=[10.^-20;-2;10.^-20;2;10.^-20];
solL=log10([10.^-20;5.7735*10.^-7;7.2169*10.^-27;1.1547*10.^-6;10.^-
21]);

end
```

Fractions continues positive – Fonction G

```
function G=PCF(Asolution,Ksolution,T,w)  %w relaxation parameter

Tn=T;Tn(Tn>=0)=0;Tp=T;Tp(Tp<0)=0;
Nc=size(Asolution,2); Ne=size(Asolution,1);
A=Asolution;A(A<=0)=NaN;a0=min(A);
 %Reactive sum
    Ap=Asolution;Ap(Ap<0)=0;
    SumR=@(X)Ap'*10.^(Ksolution+Asolution*X)+abs(Tn);
 %product sum
    An=Asolution;An(An>0)=0;
    SumP=@(X)Tp-An'*10.^(Ksolution+Asolution*X);
 %fixed point map G
G=@(X)X+w*(eye(Nc)/a0)*(log10(SumP(X))-log10(SumR(X)));

end
```

Accélération d'Anderson

```
function[X,i,t,Rh,tab,M,M1]=PCF_AA(G,x,mMax,itmax,atol,rtol,droptol,be
ta,AAstart)
% This performs fixed-point iteration with or without Anderson
% acceleration for a given fixed-point map G and initial
% approximate solution x.
%
% Required inputs:
% G = fixed-point map (function handle); form gval = g(x).
% x = initial approximate solution (column vector).
%
% Optional inputs:
% mMax = maximum number of stored residuals (non-negative integer).
% NOTE: mMax = 0 => no acceleration.
% itmax = maximum allowable number of iterations.
% atol = absolute error tolerance.
% rtol = relative error tolerance.
% droptol = tolerance for dropping stored residual vectors to improve
% conditioning: If droptol > 0, drop residuals if the
% condition number exceeds droptol; if droptol <= 0,
% do not drop residuals.
% beta = damping factor: If beta > 0 (and beta ~= 1), then the step is
% damped by beta; otherwise, the step is not damped.
% NOTE: beta can be a function handle; form beta(iter), where iter is
% the iteration number and 0 < beta(i) <= 1.
% AAstart = acceleration delay factor: If AAstart > 0, start
acceleration
% when i = AAstart.
%
% Output:
% X = approximate solutions vector at each iteration
% i = final iteration number.
```

```
% t=vecteur des temps de calcul parcourus par chaque iteration.
% Rh = residual history matrix (iteration numbers and residual norms).
% Set the method parameters.
if nargin < 2, error('AndAcc requires at least two arguments.'); end
if nargin < 3, mMax = min{10, size(x,1)}; end
if nargin < 4, itmax = 100; end
if nargin < 5, atol = 1.e-10; end
if nargin < 6, rtol = 1.e-10; end
if nargin < 7, droptol = 1.e10; end
if nargin < 8, beta = 1; end
if nargin < 9, AAstart = 0; end
tstart=tic;
t0 = cputime;
% Initialize the storage arrays.
Rh = []; % Storage of residual history.
DG = []; % Storage of g-value differences.
M=[];M1=[];tab=[];
% Initialize printing.
if mMax == 0
   fprintf('\n No acceleration');
elseif  mMax > 0
          fprintf('\n Anderson acceleration, mMax = %d \n',mMax);
     else
          error('AndAcc.m: mMax must be non-negative');
end
fprintf('\n iter res_norm \n');
% Initialize the number of stored residuals.
mAA = 0;
% Top of the iteration loop.
for i = 0:itmax
    % Apply g and compute the current residual norm.
    Gval = G(x);
    Fval = Gval - x;
    Rn = norm(Fval);
    fprintf(' %d %e \n', i, Rn);
    Rh = [Rh;[i,Rn]];
    % Set the residual tolerance on the initial iteration.
    if i == 0
       tol = max(atol,rtol*Rn);
    end
    % Test for stopping.
    if Rn <= tol
       fprintf('Terminate with residual norm = %e \n\n', Rn);
       break;
    end
    if mMax == 0 || i < AAstart
     % Without acceleration, update x <- g(x) to obtain the next
      % approximate solution.
      x = Gval;
    else
   % Apply Anderson acceleration.
   % Update the df vector and the DG array.
```

```
        if i > AAstart
           dF = Fval-F_old;
           if mAA < mMax
              DG = [DG Gval-G_old];
           else
              DG = [DG(:,2:mAA) Gval-G_old];
           end
           mAA = mAA + 1;
        end
    F_old = Fval;
    G_old = Gval;
    if mAA == 0
    % If mAA == 0, update x <- g(x) to obtain the next approximate
solution.
      x = Gval;
    else
    % If mAA > 0, solve the least-squares problem and update the
    % solution.
       if mAA == 1
       % If mAA == 1, form the initial QR decomposition.
         R(1,1) = norm(dF);
         Q = R(1,1)\dF;
      else
       % If mAA > 1, update the QR decomposition.
         if mAA > mMax
   % If the column dimension of Q is mMax, delete the first column and
             % update the decomposition.
            [Q,R] = qrdelete(Q,R,1);
             mAA = mAA - 1;
             % The following treats the qrdelete quirk described
below.
             if size(R,1) ~= size(R,2)
                Q = Q(:,1:mAA-1); R = R(1:mAA-1,:);
             end
% Explanation: If Q is not square, then qrdelete(Q,R,1) reduces the
% column dimension of Q by 1 and the column and row
% dimensions of R by 1. But if Q *is* square, then the
% column dimension of Q is not reduced and only the column
% dimension of R is reduced by one. This is to allow for
% MATLAB's default "thick" QR decomposition, which always
% produces a square Q.
         end
% Now update the QR decomposition to incorporate the new column.
     for j = 1:mAA - 1
         R(j,mAA) = Q(:,j)'*dF;
         dF = dF - R(j,mAA)*Q(:,j);
     end
     R(mAA,mAA) = norm(dF);
     Q = [Q,R(mAA,mAA)\dF];
  end
  if droptol > 0
% Drop residuals to improve conditioning if necessary.
```

```
      condDF = cond(R) ;
      M=[M;[i,cond(R)]];

      while condDF > droptol && mAA > 1
            fprintf('cond(D)=%e,reducing mAA to %d \n',condDF,mAA-1);
            [Q,R] = qrdelete(Q,R,1);
            DG = DG(:,2:mAA);
            mAA = mAA - 1;
% The following treats the qrdelete quirk described above.
            if size(R,1) ~= size(R,2)
               Q = Q(:,1:mAA); R = R(1:mAA,:);
            end
            condDF = cond(R);
      end
      M1=[M1;[i,condDF]];
  end
% Solve the least-squares problem.
   gamma = R\(Q'*Fval);
     % alpha
   Alpha(1)=gamma(1);
   for h=2:mAA
       Alpha(h)=gamma(h)-gamma(h-1);
   end
   Alpha(mAA+1)=1-gamma(mAA);
   Coefnorm=norm(Alpha,1);
   fprintf('%d %e %e %e %d \n', i, Rn, condDF, Coefnorm);
   tab=[tab;[i,Rn,condDF,Coefnorm]];
% Update the approximate solution.
   x = Gval - DG*gamma;
%Apply damping if beta is function handle or if beta>0(and beta~= 1).
   if isa(beta,'function_handle')
       x = x - (1-beta(i))*(Fval - Q*R*gamma);
   else
       if beta > 0 && beta ~= 1
          x = x - (1-beta)*(Fval - Q*R*gamma);
       end
   end
    end
 end
 x_rec(:, i+1) = real(x);
 fintime = cputime;
 telapsed = toc(tstart) ;
 t_rec(i+1) = (cputime - t0);
 display(t_rec(i+1),'CPU usage:');
end
X=x_rec;t=t_rec;
% Bottom of the iteration loop.
if Rn > tol && i == itmax
   fprintf('\n Terminate after itmax = %d iterations. \n', itmax);
   fprintf(' Residual norm = %e \n\n', Rn);
end
```

```
end
```

Méthodes MPE et RRE

Mode cyclique :

```
function[S,Rc,RESC,KOUT,t]=cycle(G,x0,N0,N,kmax,ncycle,epsc,eps,ipres,
ipres1,method)

   % x0 : INITIAL VECTOR. INPUT ARRAY OF DIMENSION NDIM.
   % s : THE FINAL APPROXIMATION PRODUCED BY THE SUBROUTINE. ARRAY OF
DIMENSION NDIM.
   % N0 : NUMBER OF ITERATIONS PERFORMED BEFORE CYCLING IS
STARTED,I.E., BEFORE MPE OR RRE IS APPLIED FOR THE FIRST TIME
   % N : NUMBER OF ITERATIONS PERFORMED BEFORE MPE OR RRE IS APPLIED
IN EACH CYCLE AFTER THE FIRST CYCLE.
   % kmax : WIDTH OF EXTRAPOLATION. ON EXIT FROM SUBROUTINE MPERRE IN
EACH CYCLE, THE ARRAY S IS, IN FACT, THE APPROXIMATION S(N0,kmax) IN
THE FIRST CYCLE,AND S(N,kmax) IN THE FOLLOWING  CYCLES.
   % ncycle: MAXIMUM NUMBER OF CYCLES ALLOWED.
   % RESC : L2-NORM OF THE RESIDUAL FOR S AT THE END OF EACH CYCLE.
RETRIEVED AT THE END OF THE NEXT CYCLE.
   % epsc : AN UPPER BOUND ON RESC/RESP, SOME RELATIVE RESIDUAL FOR S,
USED IN THE STOPPING CRITERION. HERE RESP IS THE L2-NORM OF THE
RESIDUAL FOR S(NO,KMAX) AT THE END OF THE FIRST CYCLE, I.E., ON EXIT
FROM FUNCTION MPERRE THE FIRST TIME.IF RESC.LE.EPSC*RESP AT THE END OF
SOME CYCLE, THEN ONE ADDITIONAL CYCLE IS PERFORMED, AND THE
CORRESPONDING S(N,KMAX) IS ACCEPTED AS THE FINAL APPROXIMATION, AND
THE FUNCTION IS EXITED.
   %t: CPU TIME
   %method : 1 if MPE , 2 if RRE
t0 = cputime;
eps1=1.e-32;eps2=10.^-16;
NDIM=numel(x0);
Rc=[];
for IC=1:ncycle
    if ipres==1||ipres1==1
       fprintf('CYCLE NO. is %d,',IC)
    end
    NN=N;
    if(IC==1)
       NN=N0;
    end
    if ipres==1||ipres1==1
       fprintf('NUMBER OF ITERATIONS PRIOR TO EXTRAPOLATION IS %d,
WIDTH OF EXTRAPOLATION IS %d %n',NN,kmax)
    end
    for  J=1:NN
         y=G(x0);
         for i=1:NDIM
              x0(i)=y(i);
         end
```

```
    end
[R,Rh,S,gamma,res,res1,KOUT]=MPERRE(G,x0,kmax,eps,eps1,eps2,ipres,ipre
s1,method]
    if(IC==1)
       RESP=R(1,1);
    end
    RESC=R(1,1);
    if(RESC<=epsc*RESP)
       break;
    end
    for i=1:NDIM
        x0(i)=S(i);
    end
   t=cputime-t0;
   Rc=[Rc;Rh];
end
return ;

end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Sans mode cyclique :

```
function[R,Rh,S,gamma,res,res1,KOUT,tmperre]=MPERRE(G,x0,kmax,eps,eps1
,eps2,ipres,ipres1,method)

   %THIS FUNCTION APPLIES THE MINIMAL POLYNOMIAL EXTRAPOLATION (MPE)OR
THE REDUCED RANK EXTRAPOLATION(RRE) METHODS TO A VECTORSEQUENCE
X0,X1,X2,..., THAT IS OFTEN GENERATED BY A FIXED POINT ITERATIVE
TECHNIQUE. BOTH MPE AND RRE ARE ACCELERATION OF CONVERGENCE (OR
EXTRAPOLATION) METHODS FOR VECTOR SEQUENCES. EACH METHOD PRODUCES A
TWO-DIMENSIONAL ARRAY S(N,k) OF APPROXIMATIONS TO THE LIMIT OR
ANTILIMIT OF THE SEQUENCE IN QUESTION. THE IMPLEMENTATIONS EMPLOYED IN
THE PRESENT FUNCTION GENERATE THE SEQUENCES
S(0,0)=x0,S(0,l),S(0,2),... .
   %method: IF method=l, THEN MPE IS EMPLOYED. IF method=2, THEN RRF,
IS EMPLOYED.
  %S: THE APPROXIMATION S(0,k) PRODUCED BY THE SUBROUTINE FOR EACH k.
ON EXIT, S IS S(0,KOUT).
  %kmax: A NONNEGATIVE INTEGER.THE MAXIMUM WIDTH OF EXTRAPOLATION
ALLOWED. THUS THE NUMBER OF THE VECTORS X0,X1,X2,...,EMPLOYED IN THE
PROCESS IS kmax+2 AT MOST. I
  %KOUT: A NONNEGATIVE INTEGER. KOUT IS DETERMINED BY A SUITABLE
STOPPING CRITERION, AND DOES NOT EXCEED kmax.THE VECTORS
%ACTUALLY EMPLOYED BY THE EXTRAPOLATION PROCESS ARE
%X0,X1,X2 ,..., XP, WHERE P=KOUT+l.
  %res: AN ESTIMATE FOR L2-NORM OF TE RESIDUAL FOR A NONLINEAR SYSTEM
FOR EACH k. ON EXIT, THIS k IS KOUT.
  %res1: L2-NORM OF THE RESIDUAL ACTUALLY COMPUTED FROM S(0,k) FOR
EACH k. (THE RESIDUAL VECTOR FOR ANY VECTOR VEC IS TAKEN
%AS (F(VEC)-VEC).) ON EXIT, THIS k IS KOUT.
```

```
   %eps: AN UPPER BOUND ON res/R(0,0), THE RELATIVE RESIDUAL FOR S,
USED IN THE STOPPING CRITERION. NOTE THAT R(0,0)=L2-NORM
%OF THE RESIDUAL FOR x0, THE INITIAL VECTOR. IF, FOR SOME K,
%RES<=eps*R(0,0), THEN THE CORRESPONDING S(0,k) IS ACCEPTED AS THE
FINAL APPROXIMATION, AND THE FUNCTION IS EXITED WITH KOUT=K.
%IF S(0,KMAX) iS NEEDED, THEN eps SHOULD BE
%SET EQUAL TO ZERO.
  %ipres:IF ipres=l, THEN res IS PRINTED FOR ALL k, k=0,1,... .
%OTHEWISE, IT IS NOT.
  %ipres1: IF ipres1=1, THEN resl IS COMPUTED AND PRINTED FOR ALL
%k, k=0,1,... . OTHERWISE, IT IS NOT.
  %tmperre : CPU TIME
T0 = cputime;
n=numel(x0);
Rh=[];
if ipres==1&&ipres1==1
   fprintf('\n K RES RES1 \n')
end
Nc=size(x0,1);
y=x0;
for k=0:kmax
  %COMPUTATION OF THE VECTOR XJ, J=K+1, FROM XK, AND COMPUTATION OF UK
    z=G(y);
     for i=1:n
         y(i)=z(i)-y(i);
     end

 %DETERMINATION OF THE ORTHoNORMAL VECTOR QK FROM UK BY THE MODIFIED
     %GRAM-SCHMIDT PROCESS
     if k==0
         R(1,1)=norm(y);
         Q(:,1)=(1/R(1,1)).*y;
     else
        %%%
     for j=1:k-1
         r=0;
         for i=1:n
             r=r+Q(i,j).*y(i);
         end
         R(j,k+1)=r ;
         for i=1:n
             y(i)=y(i)-R(j,k+1)*Q(i,j);
         end
     end
     R(k+1,k+1)=norm(y);
     if R(k+1,k+1) > eps1*R(1,1) && k+1<kmax
        hp=1d0/R(k+1,k+1);
        for j=1:n
            Q(j,k+1)=hp*y(j);
        end
     else
         if R(k+1,k+1) <= eps1*R(1,1)
```

```
            EEE=eps1;
            fprintf('R(%d,%d) is less or equal than %e
*R(0,0)',k,k,EEE)
            break;

         end
         end
 end
 %END OF COMPUTATION OF THE VECTOR QK
     %COMPUTATION OF THE GAMMA'S FOR MPE
     if method==1
        %Method MPE
       %[gamma,res] = mpe(R,k);
        a=R(1:k,1:k);
        b=-R(1:k,k+1);
        c = backsub(a,b);
        c = [c; 1];
        if abs(sum(c))<= eps2
            fprintf('S(0,%d) is not defined',k)
            return
        end
        gamma = c / sum(c);
        res = abs(gamma(end)) * R(end, end);
    end
    %END OF COMPUTATION OF GAMMA'S FOR MPE

    %COMPUTATION OF GAMMA'S FOR RRE
    if method==2
       [gamma, res] = rre(R, k);
    end
    %END OF COMPUTATION OF GAMMA FOR RRE

     KOUT=k ; %#ok<*NASGU>
     if ipres==1 && ipres1~=1
        fprintf('%d %e',k,res)
     end
     if res<=eps*R(1,1) || R(k+1,k+1)<=eps1*R(1,1) || k+1==kmax ||
ipres1==1
        %COMPUTATION OF THE APPROXIMATION S(0,K)
        xi(1)=1-gamma(1);
        for j=1:k-1
            xi(j+1)=xi(j)-gamma(j+1);  %#ok<*AGROW>
        end
        for i=1:n
            S(i)=x0(i);
        end
        for j=1:k-1
            hp=0;
           for i=j:k-1
                hp=hp+R(j,i+1)*xi(i+1);
           end
           for i=1:n
```

```
            S(i)=S(i)+hp*Q(i,j);
          end
       end
    end


    %END OF COMPUTATION OF THE APPROXIMATION S(0,K)
    %EXACT COMPUTATION OF RESIDUAL L2-NORM
    if ipres1==1

      % y=eq_chim_jac4(S');
      y=G(S');
      res1=0;
      for i=1:n
          res1=res1+(y(i)-S(i))^2;
      end
      res1=sqrt(res1);
        %res1=norm(y-S);

    end


    %END OF EXACT COMPUTATION OF RESIDUAL L2-NORM
     if ipres==1

        fprintf('%d %e %e \n',k,res,res1)
        %fprintf('%d %e %e %e \n',k,res,res1)
     end
          if ipres~=1
               fprintf('%d %e \n',k,res1)
           end

     if res<=eps*R(1,1) || R(k+1,k+1)<=eps1*R(1,1)
        return;
     end
     for i=1:n
        y(i)=z(i);
     end
    % y=z;
      tmperre = cputime - T0;
     Rh=[Rh;[k,res,res1]];
end

 %display(cpu,'CPU MPERRE time (%):');


end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Reduced Rank Extrapolation
function [gamma, residual] = rre(R, k)
    e = ones(k+1, 1);
```

```
    d = backsub(R, backsub(R', e));
    lambda = 1 / sum(d);
    gamma = lambda * d;
    residual = sqrt(lambda);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Minimal Polynomial Extrapolation
function [gamma,residual] = mpe(R,k)
    m=R(1:k,1:k);
    n=-R(1:k,k+1);
    c = backsub(m,n);
    c = [c; 1];

    gamma = c / sum(c);
    residual = abs(gamma(end)) * R(end, end);
end
```

Représentation graphique de convergences

Avec la méthode AA

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Anderson Acceleration
% Pour le test d'acide gallique

tot=1.e-3 ;Al_ex=2.03*10.^-5 ; H3L_ex=2.59*10.^-7 ;
%initial guess:
i1=log10([10.^-11;5*10.^-4]);i2=log10([5.012*10.^-10;10.^-9]);
T=[1.e-3 ;1.e-3] ;
%Fixed point map
[KSOLUTION,lnK,ASOLUTION,SOLUTIONNAMES,C]=get_equilib_defn;
pH=5.8;
[Ksolution,Asolution]=get_equilib_pH(KSOLUTION,ASOLUTION,pH) ;
G=PCF(Asolution,Ksolution,T,1) ;

%Cas i1 :
[X1,I1,t1,Rh1,tab1,M1,N1]=PCF_AA(G,i1,1,200,1.e-10,1.e-10,1.e10,1,0) ;
[X2,I2,t2,Rh2,tab2,M2,N2]=PCF_AA(G,i1,2,200,1.e-10,1.e-10,1.e10,1,0) ;
[X3,I3,t3,Rh3,tab3,M3,N3]=PCF_AA(G,i1,3,200,1.e-10,1.e-10,1.e10,1,0) ;
[X4,I4,t4,Rh4,tab4,M4,N4]=PCF_AA(G,i1,4,200,1.e-10,1.e-10,1.e10,1,0) ;
[X5,I5,t5,Rh5,tab5,M5,N5]=PCF_AA(G,i1,5,200,1.e-10,1.e-10,1.e10,1,0) ;


Al1=10.^(X1(1,:));Al2=10.^(X2(1,:));Al3=10.^(X3(1,:));
Al4=10.^(X4(1,:)); Al3=10.^(X5(1,:));
H3L1= 10.^(X1(2,:));H3L2=10.^(X2(2,:));H3L3=10.^(X3(2,:));
H3L4=10.^(X4(2,:)); H3L5=10.^(X5(2,:));
```

```
IT1=[0:I1-1];IT2=[0:I2-1];IT3=[0:I3-1]; IT4=[0:I4-1] ; IT5=[0:I5-1] ;

%Cas i2 :
[X12,I12,t12,Rh12,tab12,M12,N12]=PCF_AA(G,i2,1,200,1.e-10,1.e-
10,1.e10,1,0) ;
[X12r,I12r,t12r,Rh12r,tab12r]=PCF_AA(G,i2,1,200,1.e-10,1.e-
10,1.e10,1,0) ; %avec terme relaxation w=0.3
[X22,I22,t22,Rh22,tab22,M22,N22]=PCF_AA(G,i2,2,200,1.e-10,1.e-
10,1.e10,1,0) ;
[X32,I32,t32,Rh32,tab32,M32,N32]=PCF_AA(G,i2,3,200,1.e-10,1.e-
10,1.e10,1,0) ;
[X42,I42,t42,Rh42,tab42,M42,N42]=PCF_AA(G,i2,4,200,1.e-10,1.e-
10,1.e10,1,0) ;
[X52,I52,t52,Rh52,tab52,M52,N52]=PCF_AA(G,i2,5,200,1.e-10,1.e-
10,1.e10,1,0) ;


Al12=10.^(X12(1,:));Al22=10.^(X22(1,:));Al32=10.^(X32(1,:));
Al42=10.^(X42(1,:)); Al32=10.^(X52(1,:));
H3L12= 10.^(X12(2,:));H3L22=10.^(X22(2,:));H3L32=10.^(X32(2,:));
H3L42=10.^(X42(2,:)); H3L52=10.^(X52(2,:));


IT12=[0:I12-1];IT22=[0:I22-1];IT32=[0:I32-1]; IT42=[0:I42-1] ;
IT52=[0:I52-1] ;

% plot convergence to the solution
subplot(221) % cas1-Al
semilogy([0 60],[Al_ex Al_ex], 'r^-',[0 60],[tot tot],'k^-',IT1,Al1,
'r',IT2,Al2, 'b',IT3,Al3, 'g',IT4,Al4, 'v',IT5,Al5, 'k')
legend('Solution Al3+','max(Al3+)','m=1','m=2','m=3','m=4','m=5')
subplot(222) % cas1-H3l
semilogy([0 120],[H3L_ex H3L_ex], 'r^-',[0 60],[tot tot],'k^-
',IT1,H3L1, 'r',IT2,H3L2, 'b',IT3,H3L3, 'g',IT4,H3L4, 'v',IT5,H3L5,
'k')
legend('Solution H3L','max(H3L)','m=1','m=2','m=3','m=4','m=5')
subplot(223) % cas2-Al
 semilogy([0 60],[Al_ex Al_ex], 'r^-',[0 60],[tot tot],'k^-
',IT12,Al12, 'r',IT22,Al22, 'b',IT32,Al32, 'g',IT42,Al42,
'v',IT52,Al52, 'k')
legend('Solution Al3+','max(Al3+)','m=1','m=2','m=3','m=4','m=5')
subplot(224) % cas2-H3l
semilogy([0 120],[H3L_ex H3L_ex], 'r^-',[0 60],[tot tot],'k^-
',IT1,H3L12, 'r',IT22,H3L22, 'b',IT32,H3L32, 'g',IT42,H3L42,
'v',IT52,H3L52, 'k')
legend('Solution H3L','max(H3L)','m=1','m=2','m=3','m=4','m=5')

%plot residual norm
% cas1
subplot(221)
semilogy(Rh1(:,1),Rh1(:,2),'r' ,Rh2(:,1),Rh2(:,2),'b')
legend(Anderson(m=1)',Anderson m>=2)')
```

```
% cas2
subplot(222)
semilogy(Rh12(:,1),Rh12(:,2),'r' ,Rh22(:,1),Rh22(:,2),'b'
,Rh12r(:,1),Rh12r(:,2) ),'g')
legend('Anderson(m=1)', 'Anderson m>=2)', 'Anderson(m=1),k=0.3')
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%plot cond(Fk)
%cas 1
subplot(221)
semilogy(M1(:,1),M1(:,2), 'r',M2(:,1),M2(:,2), 'b')
legend('Anderson(m=1)', 'Anderson m=2)')
subplot(222)
semilogy(M3(:,1),M3(:,2), 'r',N3(:,1),N3(:,2), 'b')
legend('no condition number monitoring', 'condition number
monitoring')
%cas 2
subplot(221)
semilogy(M12(:,1),M12(:,2), 'r',M22(:,1),M22(:,2), 'b')
legend('Anderson(m=1)', 'Anderson m=2)')
subplot(222)
semilogy(M32(:,1),M32(:,2), 'r',N32(:,1),N32(:,2), 'b')
legend('no condition number monitoring', 'condition number
monitoring')
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%plot CPU time (s)
semilogy(t1,H3L1,'r',t2,H3L2,'b',t3,H3L3,'g')
legend('Anderson(m=1)', 'Anderson m=2)', 'Anderson m=3)')

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```