INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 1–16

A FORMULATION FOR FULLY RESOLVED SIMULATION (FRS) OF PARTICLE-TURBULENCE INTERACTIONS IN TWO-PHASE FLOWS

SOURABH V. APTE AND NEELESH A. PATANKAR

Abstract. A numerical formulation for fully resolved simulations of freely moving rigid particles in turbulent flows is presented. This work builds upon the fictitious-domain based approach for fast computation of fluid-rigid particle motion by Sharma & Patankar ([1] Ref. J. Compt. Phy., (205), 2005). The approach avoids explicit calculation of distributed Lagrange multipliers to impose rigid body motion and reduces the computational overhead due to the particle-phase. Implementation of the numerical algorithm in co-located, finite-volume-based, energy conserving fractional-step schemes on structured, Cartesian grids is presented. The numerical approach is first validated for flow over a fixed sphere at various Reynolds numbers and flow generated by a freely falling sphere under gravity. Grid and time-step convergence studies are performed to evaluate the accuracy of the approach. Finally, simulation of 125 cubical particles in a decaying isotropic turbulent flow is performed to study the feasibility of simulations of turbulent flows in the presence of freely moving, arbitrary-shaped rigid particles.

Key Words. DNS, particle-turbulence interactions, point-particle, fully resolved particles.

1. Introduction

Many problems in nature and engineering involve two-phase flows where solid particles of arbitrary shape and sizes are dispersed in an ambient fluid (gas or liquid) undergoing time dependent and often turbulent motion. Examples include sediment transport in rivers, fluidized beds, coal-based oxy-fuel combustion chambers, biomass gasifiers, among others. These applications involve common physical phenomena, at disparate length and time scales, of mass, momentum, and energy transport across the interface between the dispersed particles and a continuum fluid.

Numerical simulations of these flows commonly employ Lagrangian description for the dispersed phase and Eulerian formulation for the carrier phase. Depending on the volumetric loading of the dispersed phase two regimes are identified: dilute ($d_p << \ell$) and dense ($d_p \approx \ell$), where d_p is the characteristic length scale of the particle (e.g. diameter), and ℓ the inter-particle distance. Furthermore, the grid resolution (Δ) used for solution of the carrier phase could be such that the particles are 'subgrid' ($d_p << \Delta$), 'partially resolved' ($d_p \sim \Delta$), or 'fully resolved' ($\Delta < d_p$). In addition, these regimes may occur in the same problem and are dependent on the particle size as well as the grid resolution. Clearly, multiscale numerical approaches are necessary to simulate various regimes of the flow.

Received by the editors September 30, 2007 and, in revised form, May 7, 2008.
2000 *Mathematics Subject Classification.* 35R35, 49J40, 60G40.

The standard approach to simulate turbulent particle-laden flows uses direct numerical simulation (DNS) [2, 3, 4], large-eddy simulation (LES) [5, 6, 7, 8] or Reynolds-Averaged Navier Stokes (RANS) approach [9] for the carrier phase whereas the motion of the dispersed phase is modeled. In all these approaches, the particles are assumed 'point-sources' compared to the grid resolutions used (so $d_p < \eta$, the Kolmogorov length scale, for DNS whereas $d_p < \Delta$, the grid size, in LES or RANS). The fluid volume displaced by the particles are presumed negligible. Recently, an improved approach, based on mixture theory and considering the volumetric displacements of the fluid by the particles was developed for DNS/LES of particle-laden flows [10]. However, both of these approaches model the interactions between the fluid and the particles, use drag and lift correlations to approximate the drag and lift forces on the particles. The accuracy of these approaches in capturing the complex particle-fluid interactions in turbulent flows depends on the validity of simplified drag and lift laws.

Fully resolved simulation (FRS) of these flows require grid resolutions finer than the characteristic size of particles. In this approach, all scales associated with the particle motion are resolved and the drag/lift forces on the particles are *directly* evaluated rather than modeled. Considerable work has been done on fully resolved simulations of particles in laminar flows. Arbitrary Lagrangian-Eulerian (ALE) method [11], distributed Lagrange multiplier/fictious domain (DLM) based methods [12, 13, 14], Lattice Boltzmann (LBM) [15], and Immersed Boundary (IBM) based methods [16, 17, 18] have been proposed and used. These methods have been applied to simulate a modest number of particles (around 1000s) at low Reynolds numbers. In spite of several different numerical schemes, full three-dimensional direct simulations of two-phase turbulent flows in realistic configurations are rare. There are only few three-dimensional turbulent flow studies in canonical configurations on fully resolved rigid particles [19, 20] . There appears to be no reported study of fully resolved moving particles in complex geometries. In the present work, a fictitious domain based approach for motion of arbitrary rigid particles is implemented in a structured finite-volume solver capable of simulating turbulent flows. The approach is based on an efficient numerical algorithm proposed by Patankar [21] to constrain the flow field inside the particle to a rigid body motion. This facilitates simulation of large-number of particles by reducing the overhead associated with the computation of particle motion.

The paper is arranged as follows. A mathematical formulation of the basic scheme is described briefly. Numerical implementation of the scheme in a co-located grid, finite volume framework is provided next. The numerical scheme is validated for flow over a fixed sphere at different Reynolds numbers and a freely falling sphere under gravity. Simulation of 125 cubic particles in an isotropic turbulent flow is then performed to show the feasibility of the approach to capture multiscale interactions between the particles and unsteady turbulent flows.

2. Mathematical Formulation

Let Γ be the computational domain which includes both the fluid (Γ_F) and the particle ($\Gamma_P(t)$) domains. Let the fluid boundary not shared with the particle be denoted by $\mathcal{B}$ and have a Dirichlet condition. For the present work, focus is placed on flows in closed domains and thus number of particles inside the computational domain remains fixed. Further evaluation of generalized boundary conditions is necessary for inflow-outflow systems where number of particles inside the domain may vary with time. For simplicity, let there be a single particle in the domain

and the body force be assumed constant so that there is no net torque acting on the particle. The basis of fictitious-domain based approach [12] is to extend the Navier-Stokes equations for fluid motion over the entire domain Γ inclusive of particle regions. The natural choice is to assume that the particle region is filled with a Newtonian *fluid* of density equal to the particle density (ρ_P) and some fluid viscosity (μ_F). Both the fluid and the particle regions will be assumed as incompressible and thus incompressibility constraint applies over the entire region. In addition, as the particles are assumed as rigid, the motion of the material inside the particle is constrained to be a rigid body motion. Several ways of obtaining the rigidity constraint have been proposed [12, 13, 21, 14]. We follow the formulation developed by Patankar [21] which is briefly described for completeness.

The momentum equation for fluid motion applicable in the entire domain Γ is given by:

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\,\mathbf{u}\right) = -\nabla p + \nabla\cdot\left(\mu_F\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right)\right) + \rho\mathbf{g} + \mathbf{f}, \tag{1}$$

where ρ is the density field, $\mathbf{u}$ the velocity vector, p the pressure, μ_F the fluid viscosity, $\mathbf{g}$ the gravitational acceleration, and $\mathbf{f}$ is an additional body force that enforces rigid body motion inside the particle region Γ_P. For direct numerical simulations of incompressible fluid as considered in this work, μ_F is the dynamic viscosity of the fluid. It is assumed to be constant and the viscous term can be simplified to $\mu_F\nabla^2\mathbf{u}$ using the incompressibility constraint. The viscosity μ_F is defined at the *cv* centers and then evaluated at the faces using simple arithmetic averages. In the case where the grid resolution is such that all scales of turbulence are not captured, turbulence closure may be obtained using large-eddy simulations (LES) or Reynolds Averaged Navier Stokes (RANS) models. The viscosity μ_F will then be replaced by the eddy viscosity as provided by closure models used in either approaches. However, it should be noted that the turbulence closure models are valid in the bulk flow (far from the boundary). Near the moving boundary, wall modeling techniques may become necessary.

The density ρ is given as:

$$\rho = \rho_F(1-\Theta_P) + \rho_P\Theta_P;\ \Theta_P = \begin{cases} 0 \text{ in } \Gamma_\mathrm{F} \\ 1 \text{ in } \Gamma_\mathrm{P} \end{cases} \tag{2}$$

where ρ_F and ρ_P are the fluid and particle densities, respectively, Θ_p is the indicator function that assumes a value of unity inside the particle region and zero outside. In general numerical implementations, the indicator function is smeared over a small region (proportional to the grid spacing) around the boundary giving a smooth variation. As the particle moves, so does the indicator function and thus $\mathrm{D}\Theta_p/\mathrm{D}t = 0$ on the particle boundary, where $\mathrm{D}/\mathrm{D}t()$ represents a material derivative. The continuity equation in Γ for this variable density Newtonian fluid is given as:

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) = 0. \tag{3}$$

Using the definition of ρ, expanding the above equation and noting than $\mathrm{D}\Theta_p/\mathrm{D}t = 0$ on the particle boundaries gives the incompressibility constraint over the entire domain Γ:

$$\nabla\cdot\mathbf{u} = 0, \tag{4}$$

In order to enforce that the material inside the particle region moves in a rigid fashion, a rigidity constraint is required so that it will lead to a non-zero forcing function $\mathbf{f}$ in the particle region. Different ways have been proposed to obtain $\mathbf{f}$.

Inside the particle region, the rigid body motion implies vanishing deformation rate tensor:

$$\left.\begin{array}{c} \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right) = \mathbf{D}[\mathbf{u}] = 0, \\ \mathbf{u} = \mathbf{u}^{RBM} = \mathbf{U} + \Omega \times \mathbf{r} \end{array}\right\} \text{ in } \Gamma_P, \tag{5}$$

where $\mathbf{U}$ and Ω are the particle translation and angular velocities and $\mathbf{r}$ is the position vector of a point inside the particle region from the particle centroid. The vanishing deformation rate tensor for rigidity constraint automatically ensures the incompressibility constraint inside the particle region. The incompressibility constraint gives rise to the scalar field (the pressure, p) in a fluid. Similarly, the tensor constraint $\mathbf{D}[\mathbf{u}] = 0$ for rigid motion gives rise to a tensor field inside the particle region [13]. Distributed Lagrange multipliers (DLM)-based approaches have been proposed to solve for the rigid body motion and impose the rigidity constraint which requires an iterative solution strategy. Patankar [21] proposed an approach that provides the rigidity constraint explicitly, thus reducing the computational cost significantly. Noting that the tensorial rigidity constraint can be reformulated to give:

$$\nabla \cdot (\mathbf{D}[\mathbf{u}]) = 0 \text{ in } \Gamma_P; \tag{6}$$

$$\mathbf{D}[\mathbf{u}] \cdot \mathbf{n} = 0 \text{ on particle boundary.} \tag{7}$$

a two-stage fractional-step algorithm can be devised to solve the coupled fluid-particle problem [21]. Knowing the solution at time level t^n the goal is to find $\mathbf{u}$ at time t^{n+1}.

(1) In this first step, the rigidity constraint force $\mathbf{f}$ in equation 1 is set to zero and the equation together with the incompressibility constraint (equation 4) is solved by strandard fractional-step schemes over the entire domain. Accordingly, a pressure Poisson equation is derived and used to project the velocity field onto an incompressible solution. The obtained velocity field is denoted as $\mathbf{u}^{n+1}$ inside the fluid domain and $\hat{\mathbf{u}}$ inside the particle region.
(2) The velocity field in the particle domain is obtained in a second step by projecting the flow field onto a rigid body motion. Inside the particle region:

$$\rho_P \left(\frac{\mathbf{u}^{n+1} - \hat{\mathbf{u}}}{\Delta t}\right) = \mathbf{f}. \tag{8}$$

To solve for $\mathbf{u}^{n+1}$ inside the particle region we require $\mathbf{f}$. Obtaining the deformation rate tensor from $\mathbf{u}^{n+1}$ given by the above equation and using the equations (6, 7) we obtain:

$$\nabla \cdot \left(\mathbf{D}[\mathbf{u}^{n+1}]\right) = \nabla \cdot \left(\mathbf{D}\left[\hat{\mathbf{u}} + \frac{\mathbf{f}\Delta t}{\rho}\right]\right) = 0; \tag{9}$$

$$\mathbf{D}[\mathbf{u}^{n+1}] \cdot \mathbf{n} = \mathbf{D}\left[\hat{\mathbf{u}} + \frac{\mathbf{f}\Delta t}{\rho}\right] \cdot \mathbf{n} = 0. \tag{10}$$

The velocity field in the particle domain involves only translation and angular velocities. Thus $\hat{\mathbf{u}}$ is split into a rigid body motion ($\mathbf{u}^{RBM} = \mathbf{U} + \Omega \times \mathbf{r}$) and residual non-rigid motion ($\mathbf{u}'$). The translational and rotational components of the rigid body motion are obtained by conserving the linear and angular momenta and

are given as:

$$M_P \mathbf{U} = \int_{\Gamma_P} \rho \hat{\mathbf{u}} d\mathbf{x}; \tag{11}$$

$$\mathbf{I}_P \Omega = \int_{\Gamma_P} \mathbf{r} \times \rho \hat{\mathbf{u}} d\mathbf{x}, \tag{12}$$

where M_P is the mass of the particle and $\mathbf{I}_P = \int_{\Gamma_P} \rho[(\mathbf{r} \cdot \mathbf{r})\mathbf{I} - \mathbf{r} \otimes \mathbf{r}]d\mathbf{x}$ is the moment of inertia tensor. Knowing $\mathbf{U}$ and Ω for each particle, the rigid body motion inside the particle region $\mathbf{u}^{RBM}$ can be calculated. The rigidity constraint force is then simply obtained as $\mathbf{f} = \rho(\mathbf{u}^{RBM} - \hat{\mathbf{u}})/\Delta t$. This sets $\mathbf{u}^{n+1} = \mathbf{u}^{RBM}$ in the particle domain. Note that the rigidity constraint is non-zero only inside the particle domain and zero everywhere else. In practice, the fluid flow near the boundary of the particle (over a length scale on the order of the grid size) is altered by the above procedure owing to the smearing of the particle boundary.

The key advantage of the above formulation is that the projection step only involves integrations in the particle domain with no iterations. A similar approach was recently proposed in a finite-element framework by Veeramani *et al.* [14].

3. Numerical Implementation

The above formulation was implemented and tested in a finite-volume method on staggered grids by Sharma & Patankar [1]. However, their work was limited to laminar flows and few number of particles. In this paper, we present the implementation of the above formulation in an energy-conserving, co-located grid finite-volume method. The original single-phase fluid flow solver is based on that developed by Mahesh *et al.* [23] on arbitrary shaped, unstructured grids. The main advantage of this single phase flow algorithm is that it is directly applicable to turbulent flows where numerical dissipation is undesirable. We use their approach on a simple, uniform Cartesian grids, however, the scheme can be readily implemented into an unstructured grid solver for complex configurations.

Figure 1a shows the schematic of variable storage in space. All variables are stored at the control volume (cv) center with the exception of the face-normal velocity u_{N}, located at the face centers. The face-normal velocity is used to enforce continuity equation. Capital letters are used to denote particle fields. We follow the collocated spatial arrangement for velocity and pressure field as has been used by [24, 25, 23, 26]. The main reason to use this arrangement as opposed to spatial-staggering is the flexibility of extending the scheme to unstructured grids and/or adaptive mesh refinement.

Figure 1b shows the arrangement of material volumes for cubic or spherical particles. Each material volume is cubic has a characteristic length (Δ_M) which can be compared with the background grid resolution Δ. Usually, $\Delta/\Delta_M \geq 2$ is used for better approximations of interpolated quantities between the background grid and the material volumes. In this work, conservative interpolation kernels similar to those used in Immersed Boundary Methods (IBM) are used for interpolations [22]. For example, computation of an Eulerian volume fraction field, $\Theta_p(\mathbf{x})$, from the material points at $\mathbf{x}_k$ is performed as:

$$\Theta_p(\mathbf{x}) = \sum_{k=1}^{N_m} V_k \delta_\Delta^3(\mathbf{x} - \mathbf{x}_k), \tag{13}$$

where $V_k = \Delta_M^3$ is the volume associated with each material point and δ_Δ^3 is a delta function used for interpolation. In this work, a three-point delta-function

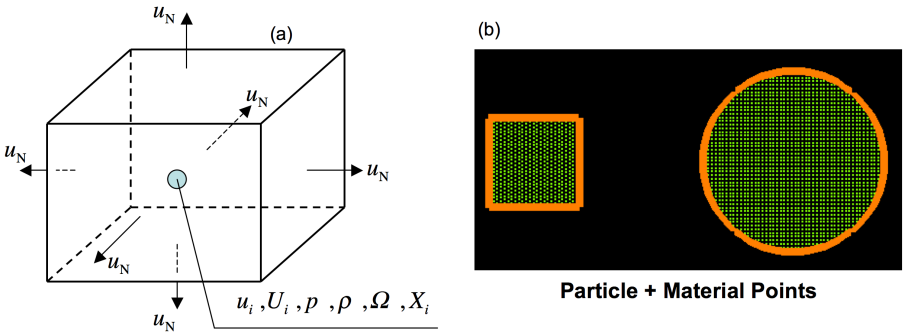


FIGURE 1. (a) Schematic of the variable storage for a co-located grid finite-volume scheme. The velocity field (u_N) represents the *face-normal* velocity and is used to enforce continuity constraint. The velocities u_i represent the Cartesian components and are co-located with the volume fraction (Θ), density (ρ), pressure (p), particle position (X_i), and the rigidity constraint force $F_{i,R}$ at the control volume (cv) centroid. In this co-located formulation, the face-normal velocity u_N is used to enforce the continuity constraint. (b) Schematic of material volume representation of particles. Each particle domain consists of cubic material volumes which retain their position with respect to the centroid of the particle. Interpolations between material volumes and background grids are used to calculate Eulerian fields such as volume fractions.

with compact support is used. In three-dimensions, the kernel is given as:

$$\delta^3_\Delta(\mathbf{x}-\mathbf{x}_k) = \frac{1}{\Delta^3}\xi(\frac{x-x_k}{\Delta})\xi(\frac{y-y_k}{\Delta})\xi(\frac{z-z_k}{\Delta}), \tag{14}$$

where the function ξ is given as:

$$\xi(r) = \begin{cases} \frac{1}{6}(5-3|r|-\sqrt{-3(1-|r|)^2+1}, & 0.5 \le |r| \le 1.5, r = \frac{(x-x_0)}{\Delta} \\ \frac{1}{3}(1+\sqrt{-3r^2+1}, & |r| \le 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

The same interpolation kernel is used to interpolate an Eulerian quantity defined at the grid centroids to the material volume centroids. The interpolation kernel is second order accurate for smoothly varying fields [22].

Assuming that the solution at time level t^n is known, the following steps are performed to advance the fluid and particle velocity fields and the particle positions to time level t^{n+1}. Integrating the governing equations over the control volume and applying Gauss' divergence theorem to convert volume integrals to surface integrals wherever possible, the discrete governing equations are derived. Accordingly, the continuity equation is

$$\frac{\rho_{cv}^{n+1}-\rho_{cv}^{n}}{\Delta t} + \frac{1}{V_{cv}} \sum_{\text{faces of cv}} \rho_{face}^{n+1} u_N^{n+1} A_{face} = 0 \tag{16}$$

where the subscript "$face$" corresponds to the face-center, A_{face} is the face area, V_{cv} the volume of the control volume, and u_N is the face-normal velocity. The density field (at any discrete time) is a linear function of the volume fraction and is given by equation (2). The particle volume fraction is a function of the position

of the particle and is obtained by interpolation procedure described above. Note that for a single phase, incompressible fluid, the above continuity equation becomes the incompressibility constraint $\sum_{\text{faces of cv}} u_{\text{N}}^{n+1} A_{face} = 0$. For the particle-laden computational domain, it also enforces the same constraint inside the fluid and particle regions. Changes in density are non-zero only near the interface, due to smearing of the particle boundary using the above interpolation kernels.

The discrete momentum equation for the i^{th} component of velocity is

$$\frac{\rho_{cv}^{n+1} u_{i,cv}^{n+1} - \rho_{cv}^{n} u_{i,cv}^{n}}{\Delta t} + \frac{1}{V_{cv}} \sum_{\text{faces of cv}} \rho_{face}^{n+1/2} u_{i,face}^{n+1/2} u_{\text{N}}^{n+1/2} A_{face} = -\frac{\partial}{\partial x_i} p_{cv}^{n+1} +$$

$$\frac{1}{V_{cv}} \sum_{\text{faces of cv}} (\tau_{ij})_{face}^{n+1/2} \, \text{N}_{j,face} A_{face} + f_{i,cv}^{n+1}, \tag{17}$$

where $(\tau_{ij})_{face}$ is the viscous stress at the faces of control volume, and $\text{N}_{j,face}$ represents the components of the outward face-normal. The velocity field $(u_{i,face})$ and the density (ρ_{face}) at the faces are obtained using arithmetic averages of the corresponding fields at two control volumes associated with the face. The interaction force f_i is used to impose the rigidity constraint within the particle domain. The above equations are solved using the fractional-step algorithm:

(1) Advance the particle positions An explicit update of the particle position is performed.

$$X_{i,P}^{n+1} = X_{i,P}^{n} + \Delta t U_{i,P}^{n+1/2}$$

An Adams-Bashforth predictor is used to obtain the particle velocity at the midpoint in time, $U_{i,p}^{n+1/2} = \frac{3}{2} U_{i,P}^{n} - \frac{1}{2} U_{i,P}^{n-1}$. For non-spherical particles, it is also necessary to advance the angular orientation of the particle about its centroid. Knowing the angular velocity at the material points $\Omega_{i,M}^{n+1/2}$, it is straight forward to obtain their new orientation with respect to the particle centroid.

(2) Evaluate Θ_P^{n+1}, ρ_{cv}^{n+1} using the equation 2

(3) Solve the momentum equations for the cell-centered velocities to obtain a predictor $(\hat{u}_{i,cv}^{n+1})$ without the rigidity constraint $(f_{i,cv}^{n+1})$.

$$\frac{\rho_{cv}^{n+1} \hat{u}_{i,cv} - \rho_{cv}^{n} u_{i,cv}^{n}}{\Delta t} + \frac{1}{V_{cv}} \sum_{\text{faces of cv}} \rho_{face}^{n+1/2} u_{i,face}^{n+1/2} u_{\text{N}}^{n+1/2} A_{face} = -\frac{\partial}{\partial x_i} p_{cv}^{n} +$$

$$\frac{1}{V_{cv}} \sum_{\text{faces of cv}} (\tau_{ij})_{face}^{n+1/2} \, \text{N}_{j,face} A_{face}$$

Arithmetic averages are used in time and space to evaluate the densities at $t^{n+1/2}$ and faces of a control volume. A Gauss-Seidel iterative solution scheme is used to solve the above non-linear partial differential equations. Second-order Adams-Bashforth predictor is used for the face-normal velocities $u_{\text{N}}^{n+1/2}$.

(4) Remove the old pressure gradient

$$\frac{\rho_{cv}^{n+1} u_{i,cv}^{*} - \rho_{cv}^{n+1} \hat{u}_{i,cv}}{\Delta t} = +\frac{\partial}{\partial x_i} (p_{cv}^{n})$$

(5) Solve the Poisson equation for pressure to impose the continuity constraint

$$\sum_{\text{faces of cv}} \frac{\delta}{\delta \text{N}} (p_{cv}^{n+1}) A_{face} = \frac{1}{\Delta t} \sum_{\text{face of cv}} \rho_{face}^{n+1} u_{\text{N}}^{*} A_{face} + V_{cv} \frac{\delta}{\Delta t} (\rho_{cv}^{n+1})$$

where $u^*_{\rm N} = 0.5(u^*_{i,nbr1} + u^*_{i,nbr2}){\rm N}_i$, is the face-normal velocity at the faces of control volumes obtained as an arithmetic average of the velocities at the neighboring control volumes ($nbr1$ and $nbr2$). The pressure Poisson equation is solved using *HYPRE*, the Algebraic Multi Grid (AMG) libraries, developed at Lawrence Livermore National Laboratory [27].

(6) Correct the velocity field by projecting out the continuity constraint. The corrected cv-center based velocity field is denoted as $u^{*n+1}_{i,cv}$ as it may not satisfy the rigidity constraint in the particle domain. This is corrected further in following steps. In a co-located grid formulation, the role of the face-normal velocity $u^{n+1}_{\rm N}$ is to enforce continuity equation. This is obtained by projecting out the face-normal pressure gradient:

$$\begin{aligned} \rho^{n+1}_{cv}\frac{(u^{*n+1}_{i,cv} - u^*_{i,cv})}{\Delta t} &= -\frac{\partial}{\partial x_i}p^{n+1}_{cv} \\ \rho^{n+1}_{face}\frac{(u^{*n+1}_{\rm N} - u^*_{\rm N})}{\Delta t} &= -\frac{\partial}{\partial {\rm N}}p^{n+1}_{cv} \end{aligned}$$

Note that it is important to properly re-construct the pressure gradient at the cv-centers $\partial p^{n+1}_{cv}/\partial x_i$. Mahesh *et al.* [23, 26] developed a face-area weighted least-squares based reconstruction that was shown to give stable and accurate results for high Reynolds number turbulent flows on arbitrary shaped, unstructured grids. We use the same algorithm here to reconstruct the pressure gradient at the cv-centers by minimizing the following expression in a least-squares sense:

$$\epsilon_{cv} = \sum_{\text{faces of cv}} \left(\frac{\partial}{\partial x_i}p^{n+1}_{cv}{\rm N}_{i,face} - \frac{\partial}{\partial {\rm N}}p^{n+1}_{cv} \right)^2 A_{face},$$

with

$$\frac{\partial}{\partial {\rm N}}p^{n+1}_{cv} = \frac{p^{n+1}_{nbr} - p^{n+1}_{cv}}{\|s_{cv,nbr}\|}$$

where $\|s_{cv,nbr}\|$ is the length of the vector connecting between the nbr and cv control volumes associated with a $face$.

(7) Evaluate the rigid body motion for each particle by calculating the translational and rotational velocity components:

$$\begin{aligned} M_P{\bf U}^{n+1}_P &= \sum_{k=1}^{M} \rho_P {\bf U}^{*n+1}_M V_k; \\ {\bf I}_P\Omega^{n+1}_P &= \sum_{k=1}^{M} ({\bf r} \times \rho_P {\bf U}^{*n+1}_M) V_k, \end{aligned}$$

where M_P and ${\bf I}_P$ are the mass and moment of inertia for the particle P, the summation is over all material points M within the particle P, $V_k (= \Delta^3_M)$ is the volume of each material point, ρ_P is the density of the particle, ${\bf U}^{*n+1}_M$ is the velocity field at the material point M obtained via interpolation from the velocity field (${\bf u}^{*n+1}_{cv}$ at the neighboring control volumes, ${\bf r}$ is the position vector of each material point from the centroid of the particle P. The interpolations are performed using the three-point delta function given by equation 14.

(8) The rigid body motion at each material point is given as ${\bf U}^{n+1,RBM}_M = {\bf U}^{n+1}_P + \Omega^{n+1}_P \times {\bf r}$. This can be interpolated to the computational grid using

the delta functions to obtain the rigid body motion $\mathbf{u}^{n+1,RBM}$ and the rigidity constraint force $f_{i,cv}^{n+1}$ at the *cv* centers:

$$\begin{aligned} f_{i,cv}^{n+1} &= \rho_{cv}^{n+1} \frac{u_{i,cv}^{RBM,n+1} - u_{i,cv}^{*n+1}}{\Delta t} \\ u_{i,cv}^{n+1} &= u_{i,cv}^{*n+1} + \frac{f_{i,cv}^{n+1} \Delta t}{\rho_{cv}^{n+1}} \end{aligned}$$

Note that the interpolations from material points to the *cv* centers provide a non-zero rigidity constraint force only in the particle domain.

This completes the advancement of the particle and velocity fields by one time-step. Extending the algorithm to multiple particles is straightforward. For multiple particles, the inter-particle collision force must be modeled to prevent the particles from penetrating each other. In case of multi-particle simulations, the accuracy of inter-particle interactions is determined by the collision model. Implementation and model details are given in detail by Glowinski *et al.* [12] and are not repeated here.

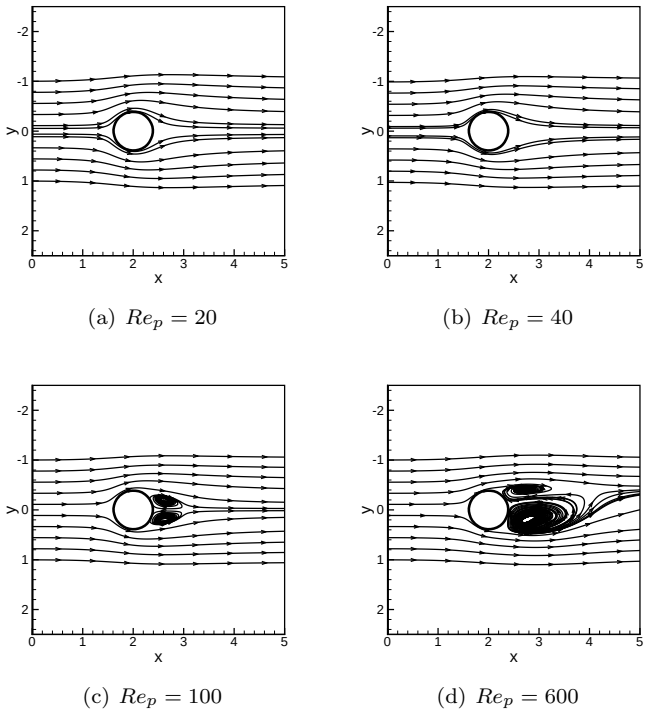
4. Results

4.1. Flow over a Fixed Sphere. Flow over a fixed sphere in a uniform stream is calculated to investigate the accuracy of the numerical scheme to predict the drag coefficient and wake effects at different Reynolds numbers. A sphere of diameter $D_p = 0.8$ units is placed at $(2, 0, 0)$ in a square domain of range $(0, -4, -4)$ to $(8, 4, 4)$ units. Uniform cubic grids (128^3) are used over the entire region. This gives around 12 grid points within the particle domain. The material volume resolution is set based on the ratio $\frac{\Delta}{\Delta_M} = 6$. Note that, in this study, the spherical particle is represented by cubic material volumes. Uniform flow of $U_\infty = 1$ units is imposed at the left boundary of the domain. A convective outflow boundary condition is imposed at the exit. Slip condition ($\frac{\partial u}{\partial \mathrm{N}} = 0$; where N represents normal to the boundary), is imposed on the boundaries in the vertical and spanwise directions. The fluid viscosity is varied to simulate flow over a sphere at different Reynolds numbers. Since the sphere is fixed, the material volumes are assigned a velocity of $\mathbf{U}_P = 0$, $\Omega_P = 0$ and setting $\mathbf{U}_M^{RBM,n+1} = 0$ in evaluating the rigidity constraint of the numerical algorithm.

Figure 2 shows the instantaneous streamlines over the fixed sphere at different Reynolds numbers. The flow remains symmetric for low Reynolds numbers ($Re_p =$ 20,40, and 100). At $Re_p = 100$, the reverse flow patterns are clearly visible. At large Reynolds number $Re_p > 300$, the flow starts to become asymmetric, shedding unsteady vortices. Table 1 compares the drag coefficients on a sphere at different Reynolds numbers with those obtained from experimental correlations [32].

TABLE 1. Comparison of drag coefficients over a fixed sphere at different Reynolds numbers.

Re_p	C_D, Present	C_D, Experiment	Error (%)
10	4.25	4.2	1.19
20	2.665	2.61	2.1
40	1.771	1.735	2.07
100	1.118	1.087	2.86
600	0.58	0.523	9.43



(a) $Re_p = 20$ (b) $Re_p = 40$

(c) $Re_p = 100$ (d) $Re_p = 600$

FIGURE 2. Instantaneous streamlines for flow over a fixed sphere at different Reynolds numbers.

The error in drag coefficient for low Reynolds numbers is less than 3%. The drag coefficient is consistently over-predicted. This may be attributed to the fact that the represented particle boundary is not perfectly spherical due to the cubical material volumes used. In addition, the boundary is smeared by the interpolation functions between the background grid and the material volumes. This has an indirect effect of making the sphere slightly larger than its actual diameter. Note that this error in spherical boundary representation can be reduced by increasing the number of material volumes (higher Δ/Δ_M ratio) or using material volumes conforming with the boundary of the sphere [31]. For large Reynolds numbers, ($Re_p = 600$) the background grid resolution of ($\Delta \approx D_P/12$) is not sufficient and affects the predicted drag coefficient. However, the algorithm is able to capture the asymmetric wake and unsteady vortex shedding as shown in Fig. 2. Accurate prediction of drag force can be obtained by further grid refinements. In most particle-laden turbulent flows, the particle Reynolds number is much lower (except for flows with high density ratios between the fluid and particle, e.g. air-particle) and a grid resolution of 10–12 points may be sufficient to capture the fluid-particle interactions.

4.2. Freely Falling Sphere. Freely falling solid sphere of density 3 kg/m^3 and diameter 0.625 m falling under gravity in a rectangular channel of cross-section 2 m $\times$ 2 m and height 8 m is simulated. The fluid viscosity is 0.05 kg/m.s and the density is 2 kg/m^3. Simulations were carried out for increasing grid refinement. The test case is same as that used by Sharma & Patankar [1] and validates our

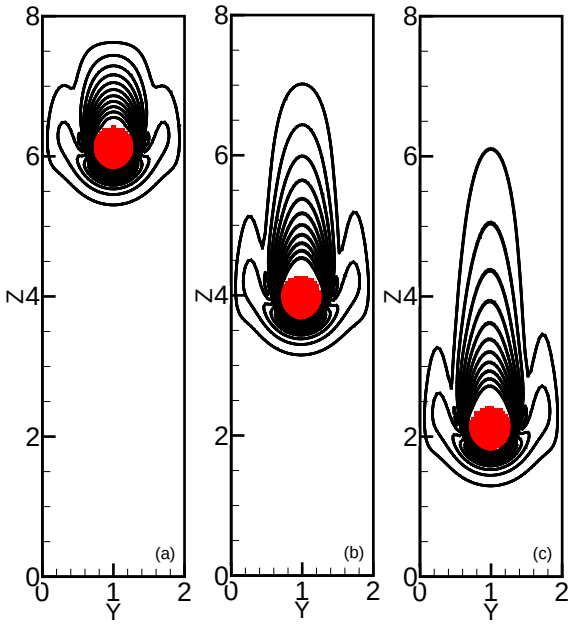
numerical implementation in a colocated finite-volume scheme. Figure 3a shows the time evolution of the fluid velocity magnitude together with the location of the sphere. As seen from the velocity magnitudes, the adjacent walls do affect the overall drag on the particle. The blocking ratio defined as the ratio of the particle diameter to the channel cross-sectional length is 0.3125. The terminal velocity of the sphere considering the wall effects can be estimated using the correlations proposed in the literature [29, 28]. Accordingly, for the present case the terminal velocity is 1.2722 m/s.

Figure 3b shows the particle terminal velocity with grid refinement. Two different studies are considered: (a) keeping the Courant-Friedrichs-Lewis (CFL = 0.1) number fixed and (b) keeping the time step fixed ($\Delta t = 2.5 \times 10^{-3}$) with grid refinement. Five different grid resolutions (Δ = 1/25,1/30,1/40,1/50,1/60) with cubical elements are used. The resolution of each material volume is fixed such that $\Delta/\Delta_M = 2$. The terminal velocity obtained with fixed CFL number is 1.288 whereas that with fixed time step is approximately 1.295. With fixed time-steps under grid refinement, the CFL number increases as the particle approaches its terminal velocity. Increased CFL numbers may lead to decreased accuracy over longer periods of time. However, with small CFL numbers, the numerical algorithm is able to predict the particle terminal velocity with good accuracy.

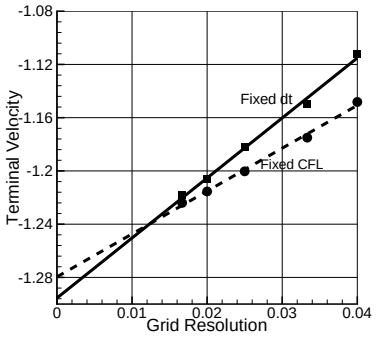
Figure 3c shows the time evolution of the particle velocity for CFL = 0.1 and grid resolution of $\Delta = 1/60$. The particle velocity is normalized by its terminal velocity and the time is normalized by the time it takes to reach 95% of the terminal velocity (t_{95}). This time evolution of the particle velocity is in qualitative agreement with the temporal behavior of sedimenting particle as observed by Mordant and Pinton [30]. The experiments were carried out at small blockage ratio (thus negligible wall effects) and thus quantitative comparison is not performed.

4.3. Particle Laden Isotropic Turbulent Flow. The numerical scheme is used to simulate particle-laden homogeneous, isotropic turbulent flow in a periodic box of length π with grid resolution of 128^3. A stationary isotropic turbulent flow is first developed using linear forcing in the fluid momentum equations proportional to the local velocity [33]. The turbulence parameters correspond to the Reynolds number of 54 based on the Taylor microscale. The turbulence intensity is $U' = 0.84$, the dissipation rate $\epsilon = 0.2$, the fluid density $\rho = 1$, and the kinematic viscosity $\nu = 0.013$. This gives the Kolmogorov length scale of $\eta = 0.056$, the Kolmogorov time scale is $\tau_K = 0.25$ the Taylor microscale $\lambda = 0.81$, the integral length scale of $L = 1.65$, the integral time scale $T = 1.98$, and $k_{max}\eta$ (the measure of resolution) is 2.28. The time-step used is $\Delta t = 1 \times 10^{-3}$.

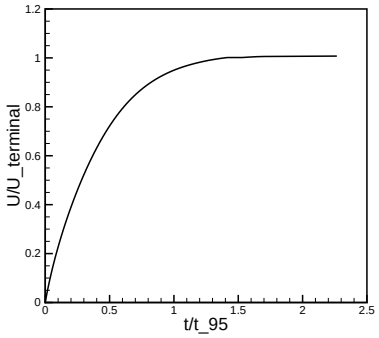
Once a stationary state is obtained, the forcing function is turned off and 125 solid cubic particles are injected into the domain, with initial uniform distribution. The length of the particles is 0.2 providing around 8 grid points over the particle domain. The particles are arranged such that they have a separation distance of $\frac{\pi}{5}$ between their nearest neighbors. The material volume resolution is based on the ratio $\frac{\Delta}{\Delta_M} = 3$. The particle density is $\rho_p = 9$ and the particle relaxation time is $\tau_p = \frac{1}{18}\frac{\rho_p}{\rho}\frac{\sigma^2}{\nu} \approx 0.9$, where σ is the characteristic length of the particle, and ν is the fluid viscosity. For the present case, the size of the particle is larger than the Kolmogorov length scale. Accordingly, different time scales can be used to normalize the particle relaxation time and define the Stokes number. Based on the Kolmogorov time-scale, the Stokes number, $St = \tau_p/\tau_K = 3.6$. The particle Reynolds number ($Re_p = \rho d_p |\mathbf{u}_{rel}|/\mu$), where $\mathbf{u}_{rel}$ is the relative velocity between



(a)



(b)



(c)

FIGURE 3. Simulation of a freely falling sphere under gravity: (a) time evolution of velocity magnitudes, (b) convergence study under grid refinement, and (c) time evolution of the particle velocity normalized by its terminal velocity. t_{95} corresponds to the particle response time or the time it takes for the particle to reach 95% of its terminal velocity.

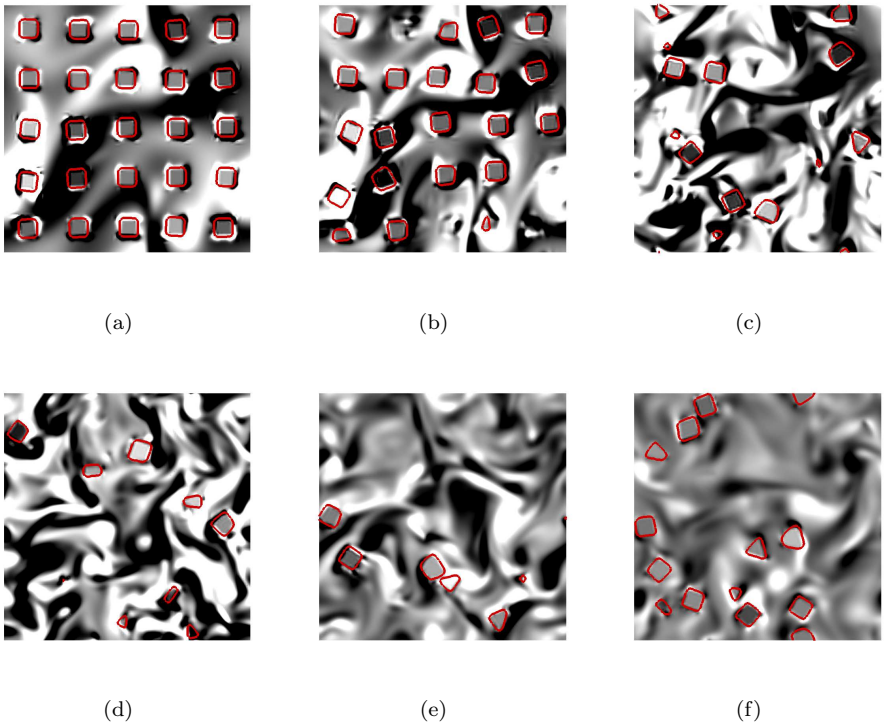
the fluid and the particle, is on the order of 20–30 in these simulations. As was shown earlier for the flow over a fixed spherical particle, the resolution of 8–12 grid cells inside the particle domain is sufficient to resolve the fluid-particle interactions in this regime. As the turbulent flow decays, the particles first accelerate, reach a maximum velocity and then decelerate.

Figure 4 shows the time-evolution of the out-of-plane vorticity contours together with the location of the particles in the symmetry plane $z = 0$. Note that since a planar cut of instantaneous particle positions is shown, the particle boundary (as seen by red lines) may not appear as a square depending upon the instantaneous location of the particle centers. Accordingly, the shapes and sizes of the particles shown in the figures appear different. However, this is just the plotting artifact and in the simulations the particles retain their size. The particles cluster in low vorticity regions as they evolve from uniform distribution. Similar results have been reported for spherical particles in isotropic turbulent flow using the Lattice-Boltzmann approach [20]. The interactions between turbulence and particle motion is fully resolved and the numerical approach can be used for further investigations of particle-laden turbulent flows.

The simulation was performed on 64-processors (IBM machines at San Diego Supercomputing Center). It requires approximately 6 seconds per time-step for this simulation. The overhead of the computing time due to computation of the rigidity constraint, the motion of particles and inter-particle collisions is only 20% of the overall time. For the present case, since the particles are initially uniformly distributed, load balancing was not an issue. The grid was partitioned such that each processor has approximately the same number of grid points. However, for inhomogeneous distribution of large number of particles, such partitioning may result in parallel load imbalance and advanced domain decomposition concepts are necessary for improved computation. In addition, in the current approach, the material volumes are present over the entire region of the particle. This is done for simplicity in the implementation of the integrations over the particle domain and also characterization of the boundary of the particle. Use of material volumes in a small band around the particles (similar to the particle level set methods) to characterize the particle-boundaries, is possible and will reduce the number of material volumes required per particle. The integration over the particle domain then can be performed using the background grid and interpolation operations.

5. Discussion

A numerical formulation for fully resolved simulations of freely moving rigid particles in turbulent flows is developed based on a co-located grid, finite-volume method. In this fictitious domain based approach, the entire computational domain is first treated as a fluid of density corresponding to the fluid or particle densities in their respective regions. The incompressibility and rigidity constraints are applied to the fluid and particle regions, respectively, by using a fractional step algorithm. The approach extends the formulation developed by Patankar [21, 1] to obtain the rigid body motion without requiring any iterative procedures. Use of consistent interpolations between the particle material volumes and the background grid and parallel implementation of the algorithm facilitates accurate and efficient simulations of large number of particles. Implementation of this approach in finite-volume based, conservative numerical solvers is presented. The numerical approach is validated for flow over a fixed sphere at various Reynolds numbers and flow generated by a freely falling sphere under gravity to show good predictive capability. Finally, simulation of 125 cubical particles in a decaying isotropic turbulent flow is performed to study the feasibility of simulations of turbulent flows in the presence of freely moving, arbitrary-shaped rigid particles. The overhead due to presence of particles and computing their motion is small and the computational speed is governed by the pressure Poisson equation used to impose the incompressibility



(a) (b) (c)

(d) (e) (f)

FIGURE 4. Temporal evolution of cubical particles in a decaying isotropic turbulent flow. The simulations are performed on a 128^3 grid and contains 125 cubical particles initially uniformly spaced. A planar cut in the $z = 0$ plane is shown, the contours indicate out-of-plane vorticity and red lines show particle shapes. Due to the planar cuts, the particle shapes *appear* different as the cubical particles move in and out of plane in this turbulent flow. The simulation was performed on 64 processors at San Diego Supercomputing Center.

constraint, making this approach attractive for large-scale simulations resolving the multiscale interactions between particles and turbulent flow.

Acknowledgments

The computing time at San Diego Supercomputing Center's Datastar machine is highly appreciated. SVA acknowledges Dr. Ki-Han Kim of the Office of Naval Research for the summer support provided under the ONR grant number N000140610697, which resulted in successful completion of the numerical implementation of the algorithm. Past interactions with Prof. K. Mahesh of University of Minnesota and Dr. Frank Ham of Stanford University on co-located grid, fractional step methods are also appreciated.

References

[1] SHARMA, N., & PATANKAR, N. A. 2005 A fast computation technique for the direct numerical simulation of rigid particulate flows., *J. Comp. Phy.*, **205** 439–457.

[2] Reade, W. C. & Collins, L. R. 2000 Effect of preferential concentration on turbulent collision rates. *Phys. Fluids* **12**, 2530–2540.
[3] Rouson, D. W. I. & Eaton, J. K. 2001 On the preferential concentration of solid particles in turbulent channel flow. *J. Fluid Mech.* **348**, 149–169.
[4] Xu, J., Dong, S., Maxey, M., & Karniadakis, G, 2003 Direct numerical simulation of turbulent channel flow with bubbles, Current Trends in Scientific Computing: ICM2002, **329**, 347–354.
[5] Wang, Q. & Squires, K.D., 1996 Large-eddy simulation of particle-laden turbulent channel flow, *Phy. Fluids*, **8**, 1207–1223.
[6] Apte, S. V., Mahesh, K., Moin, P., & Oefelein, J.C., 2003a, Large-eddy simulation of swirling particle-laden flows in a coaxial-jet combustor. *Int. J. Mult. Flow* **29**, 1311-1331.
[7] Segura, J. C., Eaton, J. K. & Oefelein, J. C. 2004 Predictive capabilities of particle-laden LES. Rep. No. TSD–156, Dept. of Mech. Engr., Stanford University.
[8] Moin, P., & Apte S.V., 2006, Large eddy simulation of multiphase reacting flows in complex combustors, *AIAA J.*, **44**, 698–710 (special issue on 'Combustion Modeling and LES: Development and Validation Needs for Gas Turbine Combustors).
[9] Sommerfeld, M., Ando, A., & Qiu, H. H. 1992 Swirling, particle-laden flows through a pipe expansion *J. Fluids. Engr.* **114**, 648–656.
[10] Apte, S.V., Mahesh, K., & Lundgren T., 2007 Accounting for Finite-Size Effects in Simulations of Two-Phase Flows *Int. J. Multiphase Flow* accepted for publication.
[11] Hu, H. H., Patankar, N.A., Zhu, M.Y., 2001, Direct numerical simulation of fluid-solid systems using the arbitrary Lagrangian-Eulerian technique, *J. Comp. Phy.*, **169**, 427–462.
[12] Glowinski, R. Pan, T.W., Hesla, T.I., Joseph, D.D., Periaux, J., 2001, A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: Application to particulate flow, *J. Comput. Phys.*, **169**, 363–426.
[13] Patankar, N., Singh, P., Joseph, D., Glowinski, R., & Pan, T., 2000 A new formulation of the distributed Lagrange multiplier/fictitious domain method for particulate flows *Int. J. Multiphase Flow* **26**, 1509.
[14] Veeramani, C., Minev, P.D., & Nandakumar, K,, 2007, A fictitious domain formulation for flows with rigid particles: A non-Lagrange multiplier version, *J. Comp. Phy.*, **224**, 867–879.
[15] Ladd, A.J.C. & Verberg, R., 2001 Lattice-Boltzmann simulations of particle-fluid suspensions, J. Statist. Phys., **104**, 1191–1251.
[16] Zhang, Z. & Prosperetti, A., 2005 A second-order method for three-dimensional particle flow simulations, *J. Comput. Phys*, **210**, 292–324.
[17] Feng, Z-G, & Michaelides, E.E., 2005, Proteus: a direct forcing method in the simulations of particulate flows, *J. Comp. Phy.*, **202**, 20–51.
[18] Taira K., & Colonius, T., 2007, The immersed boundary method: A projection approach, *J. Comp. Phy.*, **225** (2), 2118–2137.
[19] Kajishima, T., Takiguchi, S., & Miyake, Y., 1999, Modulation and subgrid scale modeling of gas-particle turbulent flow, in: D. Knight, L. Sakell (Eds.), Recent Advances in DNS and LES, Kluwer Academic Publishers, 235–244.
[20] Ten Cate, A., Derksen, J.J., Portela L.M., & van den Akker, H. E.A., 2004, Fully resolved simulations of colloiding monodisperse spheres in forced isotropic turbulence, *J. Fluid Mech.*, **519**, 233–271.
[21] Patankar, N. A. 2001 A formulation for fast computations of rigid particulate flows. *Annual Research Briefs Center Turbul. Res.*, 185–196.
[22] Peskin, C.S, 2002, The immersed boundary method, *Acta Numerica*, **11** 479–517.
[23] Mahesh, K., Constantinescu, G., and Moin, P., 2004, A new time-accurate finite-volume fractional-step algorithm for prediction of turbulent flows on unstructured hybrid meshes, *J. Comp. Phys.*, **197**, 215–240.
[24] Kim, D., & Choi, H, 2000, A Second-Order Time-Accurate Finite Volume Method for Unsteady Incompressible Flow on Hybrid Unstructured Grids, *J. Comp. Phy.*, **162**, 411–428.
[25] Ham, F., Apte, S.V., Iaccarino, G., Wu, X., Herrmann, M., Constantinescu, G., Mahesh, K., and Moin, P., 2003, Unstructured LES of reacting multiphase flows in realistic gas-turbine combustors, *Annual Research Briefs*, Center for Turbulence Research, Stanford University, 13–160, (http://ctr.stanford.edu).
[26] Mahesh, K., Constantinescu, G., Apte, S.V., Iaccarino, G., Ham, F., and Moin, P., 2006 Large-eddy simulation of reacting turbulent flows in complex geometries, *ASME J. App. Mech.*, accepted for publication.

[27] Lawrence and Livermore and National and Laboratory's and CASC and Team 2006, Scalable linear solvers, HYPRE, *http://www.llnl.gov/CASC/linear_solvers, hypre2.0.0.*
[28] Brown, P.P. & Lawler, D.F, 2003, Sphere drag and settling velocity revisited, *J. Environ. Engng.* **129**, 222–231.
[29] Fidleris,V., & Whitmore, R.L. 1961 Experimental determination of the wall effect for spheres falling axially in cylindrical vessels, *Br. J. Appl. Phys.* **12**, 490–494.
[30] Mordant, N. & Pinton, J.F., 2000, Velocity measurement of a settling sphere, *Eur. Phys. J. B* **18** 343–352.
[31] Yu, Z., Shao, X., & Wachs, A., 2006, A fictitious domain method for particulate flows with heat transfer, *J. Comp. Phy.*, **217**, 424–452.
[32] White F. M., 1991, Viscous fluid flow, McGraw-Hill, New York.
[33] Lundgren, T., 2003, Linearly forced isotropic turbulence, *Annual Research Briefs*, Center for Turbulence Research, Stanford University, 461–473, (http://ctr.stanford.edu).

Department of Mechanical Engineering, Oregon State University, Corvallis, OR 97331, USA
E-mail: sva@engr.orst.edu
URL: http://me.oregonstate.edu/people/faculty/therm_fluid/apte.html

Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA
E-mail: n-patankar@northwestern.edu
URL: http://www.mech.northwestern.edu/web/people/faculty/patankar.htm

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 17–35

MULTISCALE FEATURE DETECTION IN UNSTEADY SEPARATED FLOWS

GUONING CHEN, ZHONGZANG LIN, DANIEL MORSE, STEPHEN SNIDER,
SOURABH APTE, JAMES LIBURDY, AND EUGENE ZHANG

Abstract. Very complex flow structures occur during separation that can appear in a wide variety of applications involving flow over a bluff body. This study examines the ability to detect the dynamic interactions of vortical structures generated from a Helmholtz instability caused by separation over bluff bodies at large Reynolds number of approximately 10^4 based on a cross stream characteristic length of the geometry. Accordingly, two configurations, a thin airfoil with flow at an angle of attack of 20^0 and a square cylinder with normally incident flow are examined. A time-resolved, three-component PIV data set is collected in a symmetry plane for the airfoil, whereas direct numerical simulations are used to obtain flow over the square cylinder. The experimental data consists of the velocity field, whereas simulations provide both velocity and pressure-gradient fields. Two different approaches analyzing vector field and tensor field topologies are considered to identify vortical structures and local, swirl regions. The vector field topology uses (1) the Γ function that maps the degree of rotation rate (or pressure-gradients) to identify local swirl regions, and (2) Entity Connection Graph (ECG) that combines the Conley theory and Morse decomposition to identify vector field topology consisting of fixed points (sources, sinks, saddles) and periodic orbits, together with separatrices (links connecting them). The tensor field feature uses (1) the λ_2 method that examines the gradient fields of velocity or pressure-gradient to identify local regions of pressure minima, and (2) tensor field feature that decomposes the velocity-gradient or pressure Hessian tensor into isotropic scaling, rotation, and anisotropic stretching parts to identify regions of high swirl. The vector-field topology requires spatial integration of the velocity or pressure-gradient fields and represents a global descriptor of vortical structures. The tensor field feature, on the other hand, is based on gradients of the velocity of pressure-gradient vectors and represents a local descriptor. A detailed comparison of these techniques is performed by applying them to velocity or pressure-based data and using spatial filtered data sets to identify the multiscale features of the flow. It is shown that various techniques provide useful information about the flow field at different scales that can be used for further analysis of many fluid engineering problems of practical interest.

Key Words. Vortex detection, separated flows, multiscale feature detection, turbulence, LES/DNS, vector field topology, tensor field feature.

1. Introduction

The ability to detect discrete flow structures in fluid flow environments is of growing interest to a wide variety of applications. For instance, large scale flow

structures such as swirling, high shear rates regions and vortical structures are thought to be controlling mechanisms for chaotic mixing, unsteady pressure fields that influence fluid-surface interactions, transport in multiphase flows, and a host of other applications. A robust means of developing an understanding of how these flow structures develop, evolve, decay, and interact is of fundamental importance. To achieve this goal there needs to be a quantitative measure of the relevant flow structures. This quantitative measure should also allow for spatial distinction among structures and a means of tracking such structures in the space and time domains. Since there may be many different views on what is a flow structure, there is a wide range of defining conditions for said structures. This results in a number of possible ways of detecting the desired flow structure. The unifying requirement of the detection schemes is that they provide a quantitative measure in a complex flow environment that defines the extent of the structure elements with an acceptable spatial and temporal resolution.

In this study the goal is to identify flow structures that are generated as a result of flow separation that occurs during flow over a bluff body. Such flow separation is indicative of a Kelvin-Helmhotz shearing instability [1, 2, 3] which results in a roll-up along a highly concentrated vortex sheet (or high shear region). Flows of this nature are extremely important in determining the dynamic loading on structures, in aerodynamic flight conditions, and drag forces on man-made vehicles or animals in motion. Presented are results for two such bodies, a thin airfoil at a high angle of attack (angle between the airfoil chord and flight direction is large causing leading edge flow separation) and a square cross section object with separation at both the front and trailing edges. The flow patterns associated with both bodies are illustrated later in this paper, but the common element of concern for these flows is that the flow separation generates large swirling flow structures that are convected downstream as they change in size, shape and intensity.

2. Related Work

Traditionally, flow analysis involving turbulence and unsteady coherent structures that may be imbedded within the broad spectrum of turbulence has been based on collecting one-point and two-point statistics. However, there is a large and growing literature on swirl and vortical flow detection methods [4, 5, 6, 7]. Proper Orthogonal Decomposition (POD) [6], the λ_2 (second eigenvalue) method [4, 8], and the Γ function [9, 7], among others, have been proposed and typically used for flow analysis. Specific identification of vortex structures (or pressure minima) [4, 8, 9] and correlating vortex shedding to leading edge separation [3], have been applied. In addition to these, novel approaches developed in the scientific visualization community based on vector and tensor field visualization and topology extraction provide an alternative means to extract flow structure features [10, 11].

Recent advances in vector field topology focus on features such as fixed points, periodic orbits, and separatrices [12, 13, 14, 15, 16] in two-dimensions, which have been extended to three-dimensional steady state [17, 18, 19], and time-dependent flows [20, 21, 22, 23], respectively. To address noise in the data sets, various flow simplification algorithms have been proposed that are either topology-based [14, 24, 16] or purely geometric [25]. Symmetric tensor field analysis has also been well investigated in two-dimensions [26]. The basic constituents of tensor field topology, the wedges and trisectors have been identified in 2D, symmetric, second-order tensors. By tracking their evolution over time, these features can be combined to form more familiar field singularities (i.e. fixed points) such as saddles, nodes,

centers, or foci [26]. This work has been extended to three-dimensions [27, 28, 29] and to time-varying tensor fields [30]. Tensor field simplification techniques have also been developed [31, 32]. Analysis of asymmetric tensor fields such as the velocity gradient has been performed [33, 34]. Zhang et al. [34] propose to perform topological analysis on the eigenvalues and eigenvectors of the velocity gradient to explore flow features such as regions of compression, dilation, rotation, and stretching, which leads to *tensor field feature* extraction. In this work, the vector field topology and tensor field feature extraction techniques will be applied to flow data sets and compared with the traditional approaches based on the Γ and λ_2 methods. Furthermore, comparison of velocity-based and pressure-gradient based data and application of various feature extraction methods is performed to illustrate the potential of each technique when applied to different data sets.

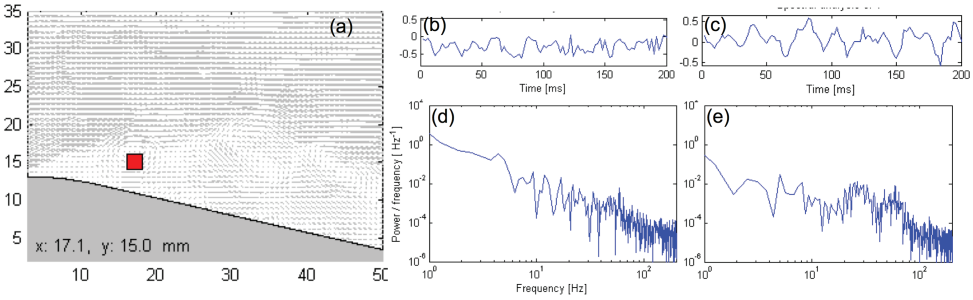


FIGURE 1. Time-resolved PIV data in the symmetry plane obtained at the OSU wind tunnel. (a) A close up view of the flow separation near the leading edge at 20^0 angle of attack, (b-c) time history of axial and vertical velocity signals, respectively, at x = 17.1 and y = 15 mm, (d-e) the corresponding power density spectra showing a broadband spectrum and time scales.

In this study two data sets have been selected to explore the capabilities of flow structure detection during separation, shown in Figures 1–2. Figure 1 shows experimentally obtained velocity components in a two dimensional plane along the centerline of a wing in a moderate Reynolds number ($Re = 6 \times 10^4$ based on the chord length). This data set was obtained using particle image velocimetry and represents a snapshot of the velocity field with a vector resolution of 0.684mm in a total field of field of 54 mm$\times$47 mm. The wing is at a 20^0 angle of attack (chord line relative to flow direction) and as such experiences a leading edge separation. The flow structures developing from this separation are of interest as they are convected downstream. The energy spectrum associated with the leading edge region shows a broadband spectrum and is typical of these separated flows.

Figure 2 shows a snapshot of a computational simulation of flow over a square cylinder at $Re \approx 10,000$ based on the inlet velocity ($U_\infty = 1.5$ m/s), the cube size (L = 0.1 m and the fluid kinematic viscosity ($\nu = 15 \times 10^{-6}$ m^2/s). A three-dimensional simulation is performed with Dirichlet conditions at the inlet, a slip condition at the top and bottom surfaces, periodic conditions in the spanwise direction, and a convective boundary condition for the outlet. The direct numerical simulation is performed based on a colocated grid, fractional step algorithm [35, 36, 37] to collect the velocity and pressure data in space and time. The flow solver has been validated with available experimental data for a variety of flow configurations involving separated turbulent flows and swirling regions [38, 39, 36, 37]. Also shown

in Figure 2 are the variations of mean axial velocity and turbulent kinetic energy in the vertical directions at three different sections. The flow separates at the leading edge corners and forms an oscillatory wake downstream giving rise to large scale vortical structures containing large levels of turbulent kinetic energy.

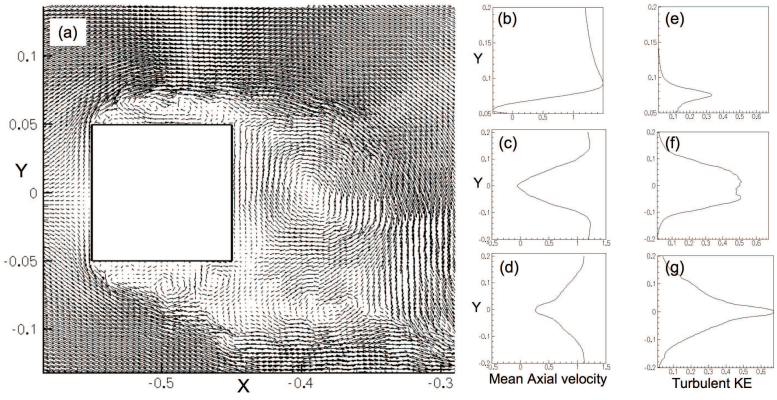


FIGURE 2. Direct numerical simulation of flow over a square cylinder at $Re = 10,000$. (a) the velocity vector field in the symmetry plane, (b) mean axial velocity variation in y-direction at midpoint of the top surface of the cylinder (c), one-length and (d) two-length downstream from the trailing edge, (e-g) the turbulent kinetic energy variations in the y-direction at corresponding sections.

These two data sets were selected for a number of reasons. They both represent leading edge flow separation with significant vortical flow structure development. Both flows are at a reasonably high Reynolds number to assure a range of scales of motion and energy. Consequently, the robust nature of the vortex detection scheme can be evaluated for these multiscale flow fields. In addition, in order to determine the ability to extract those vortical scales within a specified size or energy range a high, low or band pass filtering can be applied to the data set. Here we illustrate this by using a Gaussian filter to generate both a high and low pass filtered data set which is then analyzed using the various detection schemes. Also, there is a fundamental difference between experimentally and computationally obtained data sets. In general the computational simulation will contain both velocity and pressure field results over the full extent of the field of interest, while the experimental set will be limited to a velocity vector field (usually in two dimensions, and rarely more than two components). Consequently, experimental data sets lack the ability to use the pressure and/or the full dimensional field and its possible gradients as an indicator variable, or feature descriptor. As discussed below, several indicator variables are explored in this paper to assess the ability and distinctions of different variables to detect vortical flow structures.

The paper is arranged as follows. In the following section, the various flow analysis techniques are described in detail. These include both the vector field (velocity or pressure gradient) and tensor field (velocity gradient or Hessian of pressure) analyses. These techniques are then applied to the airfoil and square cylinder data sets and compared to assess the similarities and differences of all detection schemes in identifying vortical flow features. In addition, flow structures associated with scale separation are investigated by applying high and low-pass

filters to the original data sets. The computational data set is used to compare velocity-based and pressure-gradient based data.

3. Flow Analysis Techniques

In this section, we describe a number of existing flow descriptors and compare their effectiveness in multi-scale analysis of flows with separation and vortices.

These descriptors can be roughly divided into two categories: global and local, based on the techniques used to compute them. The *global descriptors* include the Γ-*function* and *Entity Connection Graph* (ECG) [16], both of which are derived from a vector field (velocity or pressure gradient) and requires integration over a region surrounding the point of interest. In contrast, the *local descriptors*, such as the λ_2 method and eigenvalue and eigenvector topology [34], are based on the gradient of a vector field.

3.1. Global Descriptors. In the following subsections, global descriptors based on a vector field such as the velocity or pressure-gradient are described.

3.1.1. The Γ-Function. A Γ function [9] has been proposed as a swirl strength parameter and used by one of the co-authors [7] to study pulsed jet in crossbow. This method is based on a direct measure of the local swirl tendency of the flow field by calculating the vector orientation of the feature descriptor relative to a local radius vector at a given point within the flow field. Using the velocity vector $\overline{U}_M$ in the $\mathrm{x}-\mathrm{y}$ plane as the feature descriptor the swirl strength, Γ, is determined within a local grid area A_M by:

$$\Gamma(x,y) = \frac{1}{A_M}\int_{A_M} \frac{\left(\overline{PM}\times\overline{U}_M\right)\cdot\widehat{Z}dA}{\left(\|\overline{PM}\|\|\overline{U}_M\|\right)} \tag{1}$$

where $\widehat{Z}$ is a unit vector pointing out of the (x,y)-plane, and $\overline{PM}$ is the position vector of point M within the integration stencil and the point P. This is equivalent to the summation of the *sine* of the angle between the velocity vector at points within the area A_M and the position vector from these points to the position (x,y). Consequently, it is a measure of the local swirl strength filtered by the selection of the area A_M. Because of the local normalization, the swirl can be detected within regions of large dynamic range of velocity, which is advantageous in a separated flow region.

Note that the traditional definition of Γ function is based on the velocity vector. In order to define a similar feature detector based on the pressure-gradient vector, a new function denoted as Γ_p is defined as:

$$\Gamma_p(x,y) = \frac{1}{A_M}\int_{A_M} \frac{\left(\overline{PM}\times\overline{P}'_M\right)\cdot\widehat{Z}dA}{\left(\|\overline{PM}\|\|\overline{P}'_M\|\right)} \tag{2}$$

where $\overline{P}'_M = -(\nabla p)^{\perp}$ is the pressure gradient field rotated by 90^0 in the anti-clockwise direction. The pressure gradient normal to the radial vector centered at a given point within the flow is used, and is integrated about area A_m in a similar manner as shown above for the velocity vector. In this case the swirl indication is based on a local low pressure region which is scaled by the area averaged pressure gradient aligned toward a specific location within the flow. The area of integration is selected based on the spatial scale of interest.

3.1.2. Entity Connection Graph (ECG). Vector field topology in two-dimensions consists of fixed points (sources, sinks, saddles) and periodic orbits, Figure 3), together with separatrices (links connecting them). The fixed points identify specific flow features and the separatrices provide possible paths and correlations between spatially varying structures. These entities and their interconnection can be represented by a graph called *Entity Connection Graph* (*ECG*) [16]. Note that the fixed points and periodic orbits are the nodes in the ECG and separatrices are the edges. In addition, a periodic orbit can be connected directly to a source, sink, or another periodic orbit. The ECGs of vector fields (for example, velocity and pressure gradient) can be used to identify specific flow features (such as vortex centers etc.).

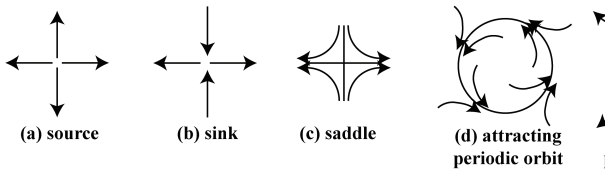


FIGURE 3. Schematic of vector field topology: (a) source, (b) sink, (c) saddle, (d) attracting and (e) repelling periodic orbits.

Mathematically, a vector field can be expressed in terms of a differential equation $\dot{x} = V(x)$. The set of solutions to it gives rise to a *flow* on the underlying domain M; that is a continuous function $\varphi : \mathbf{R} \times M \to M$ satisfying $\varphi(0, x) = x$, for all $x \in M$, and

$$\varphi(t, \varphi(s, x)) = \varphi(t + s, x) \tag{3}$$

for all $x \in M$ and $t, s \in \mathbf{R}$. Given $x \in M$, its *trajectory* is

$$\varphi(\mathbf{R}, x) := \cup_{t \in \mathbf{R}} \varphi(t, x). \tag{4}$$

$S \subset M$ is an *invariant set* if $\varphi(t, S) = S$ for all $t \in \mathbf{R}$. Observe that for every $x \in M$, its trajectory is an invariant set. Other simple examples of invariant sets include the following. A point $x \in M$ is a *fixed point* if $\varphi(t, x) = x$ for all $t \in \mathbf{R}$. More generally, x is a *periodic point* if there exists $T > 0$ such that $\varphi(T, x) = x$. The trajectory of a periodic point is called a *periodic orbit*.

Consideration of the important qualitative structures associated with vector fields on a surface requires familiarity with hyperbolic fixed points, period orbits and separatrices. Let x_0 be a fixed point of a vector field $\dot{x} = V(x)$; that is $V(x_0) = 0$. The linearization of V about x_0, results in a 2×2 matrix $Df(x_0)$ which has two (potentially complex) eigenvalues $\sigma_1 + i\mu_1$ and $\sigma_2 + i\mu_2$. If $\sigma_1 \neq 0 \neq \sigma_2$, then x_0 is called a *hyperbolic fixed point*. Observe that on a surface there are three types of hyperbolic fixed points: *sinks* $\sigma_1, \sigma_2 < 0$, *saddles* $\sigma_1 < 0 < \sigma_2$, and *sources* $0 < \sigma_1, \sigma_2$. Systems with invariant sets such as periodic orbits are considered and the definition of the limit of a solution with respect to time is non-trivial. The *alpha* and *omega limit sets* of $x \in M$ are

$$\alpha(x) := \cap_{t<0} \mathrm{cl}(\varphi((-\infty, t), x)), \quad \omega(x) := \cap_{t>0} \mathrm{cl}(\varphi((t, \infty), x))$$

respectively. A periodic orbit $\mathcal{O}$ is *attracting* if there exists $\epsilon > 0$ such that for every x which lies within a distance ϵ of $\mathcal{O}$, $\omega(x) = \mathcal{O}$. A *repelling* periodic orbit can be similarly defined ($\alpha(x) = \mathcal{O}$)). Finally, given a point $x_0 \in M$, its trajectory is a *separatrix* if the pair of limit sets $(\alpha(x), \omega(x))$ consist of a saddle fixed point and another object that can be a source, a sink, or a periodic orbit.

Figure 4 provides an example vector field (left). Fixed points are highlighted by colored dots (sources: green; sinks: red; saddles: blue). Periodic orbits are colored in green if repelling and in red if attracting. Separatrices that terminate in a source or a repelling periodic orbit are shown in green and those terminate in a sink or an attracting periodic orbit are colored in red.

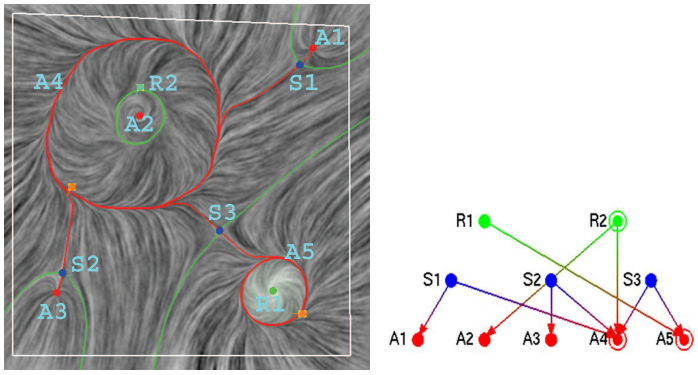


FIGURE 4. An example vector field (left) and its ECG [16] (right). The vector field contains a source (green), three sinks (red), three saddles (blue), a repelling periodic orbit (green), and two attracting periodic orbits (red). Separatrices that connect a saddle to a repeller (a source or a periodic orbit) are colored in green, and to an attractor (a sink or a periodic orbit) are colored in red. The fixed points and periodic orbits are the nodes in the ECG and separatrices are the edges.

3.2. Local Descriptors. The gradient of a vector field is an asymmetric tensor field, and the topological and geometric analysis of the vector gradient can provide additional insights to the understanding of the vector field itself. Here a well-known technique, the λ_2 method, as well as a newly developed descriptor based on the eigenvalue topology [34] are applied to the experimental and numerical data sets.

3.2.1. The λ_2 Method. The local swirl within a flow can be determined based on a local pressure minimum by assessing the gradient fields of either velocity or pressure, this is designated as the λ_2 method. Jeong and Hussain [4] provide a thorough discussion of the various criteria and argue that the Hessian of pressure be used to identify local pressure minima, and hence the vortex core.

The equations of motion for an incompressible, Newtonian fluid with constant viscosity are given by the Navier-Stokes equations:

$$\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j}(u_i u_j) = -\frac{1}{\rho}\frac{\partial P}{\partial x_i} + \nu \frac{\partial^2 u_i}{\partial x_j \partial x_j} \tag{5}$$

where u_i represents the components of the velocity vector, P the pressure field, and ν the kinematic viscosity. In addition, the velocity field must satisfy the divergence free constraint $u_{j,j} = 0$ for an incompressible fluid. Taking the gradient of the Navier-Stokes equation results in the relationship shown below between the pressure Hessian and the velocity gradient tensor separated into its symmetric and antisymmetric parts, S_{ij} and Ω_{ij}, respectively,

$$\underbrace{\left[\frac{D\mathrm{S}_{ij}}{Dt} + \mathrm{S}_{ik}\mathrm{S}_{kj} + \Omega_{ik}\Omega_{kj}\right]}_{\text{symmetric}} + \underbrace{\left[\frac{D\Omega_{ij}}{Dt} + \Omega_{ik}\mathrm{S}_{kj} + \mathrm{S}_{ik}\Omega_{kj}\right]}_{\text{antisymmetric}} = -\frac{1}{\rho}P_{,ij} + \nu u_{i,jkk}, \tag{6}$$

where $S_{ij} = (u_{i,j} + u_{j,i})/2$ and $\Omega_{ij} = (u_{i,j} - u_{j,i})/2$.

A direct measure of the local pressure minimum can be obtained by evaluation of the eigenvalues of the pressure Hessian ($P_{,ij}$). Upon ordering the eigenvalues, a positive second eigenvalue, denoted here as $\lambda_{2,p}$ expresses a local minimum. Alternatively, if the advective ($\frac{D\mathrm{S}_{ij}}{Dt}$) and viscous terms ($\nu u_{i,jkk}$) of the above gradient equation are assumed small, the strain and rotation tensors, S_{ij} and Ω_{ij}, can be used to relate the effects of the local pressure minimum. Noting that the second bracket of the equation is identically zero (it is the well-known vorticity transport equation [4]), this method examines the eigenvalues of the remaining terms on the left hand side by using the velocity gradient fields and represents an estimation of the pressure Hessian ($P_{,ij}$). This is denoted as λ_2.

The majority of the works in turbulent, separated flow use the λ_2 method mainly because the velocity field can be directly measured in laboratories and hence its gradients can be obtained. However, detailed measurement of the pressure field in a region is usually not performed. In numerical simulations, both the velocity and pressure fields are computed and allows computation of different measures of swirl strengths. By examining the results using both the λ_2 and $\lambda_{2,p}$ it is possible to assess the detection sensitivities based on the velocity-based versus pressure-gradient based fields. In this case the exclusion of the convective and viscous terms can be evaluated.

3.2.2. Tensor Field Feature: The Eigenvalue Manifold (EM). The feature of an asymmetric tensor field consists of the features of its eigenvalues and eigenvectors. In two-dimensional cases, Zhang et al. [34] define the concepts of *Eigenvalue Manifold* and *Eigenvector Manifold*, which allow the topological characterization of an arbitrary tensor field T. In the present work, we only focus on the eigenvalue feature of the velocity gradient tensor and the Hessian of the pressure. This characterization is based on the following reparameterization of the set of all 2×2 tensors,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \gamma_d \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \gamma_r \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + \gamma_s \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \tag{7}$$

where

$$\gamma_d = \frac{a+d}{2}, \quad \gamma_r = \frac{c-b}{2} \quad \gamma_s = \frac{\sqrt{(a-d)^2 + (b+c)^2}}{2} \tag{8}$$

are the *strengths* of isotropic scaling, rotation, and anisotropic stretching, respectively. For unit tensors, i.e., $\gamma_d^2 + \gamma_r^2 + \gamma_s^2 = 1$, the *eigenvalue manifold* is defined as

$$\{(\gamma_d, \gamma_r, \gamma_s) | \gamma_d^2 + \gamma_r^2 + \gamma_s^2 = 1 \text{ and } \gamma_s \geq 0\} \tag{9}$$

Notice that if the tensor T is the velocity gradient $u_{i,j}$, then the strengths of the tensor fields become:

$$\begin{aligned} \gamma_d &= \frac{1}{2}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right);\ \gamma_r = \frac{1}{2}\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right) \\ \gamma_s &= \frac{1}{2}\sqrt{\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)^2}, \end{aligned}$$

where u and v are the velocity components in the x and y directions, respectively. In a two-dimensional flow, γ_d is identically zero for an incompressible fluid due to the divergence-free constraint on the velocity field, whereas for a three-dimensional flow $\gamma_d = -1/2\,(\partial w/\partial z)$ and represents the mass flux in the z direction. The positive and negative values of the isotropic scaling thus represent expansion and contraction of fluid elements in the other direction. Similarly, γ_r represents the vorticity in the z-direction. The anisotropic stretching strength γ_s represents the rate of angular deformation and is related to regions of high shear. By decomposing the tensor into isotropic scaling, rotation, and anisotropic stretching parts, it is possible to identify regions of high swirl.

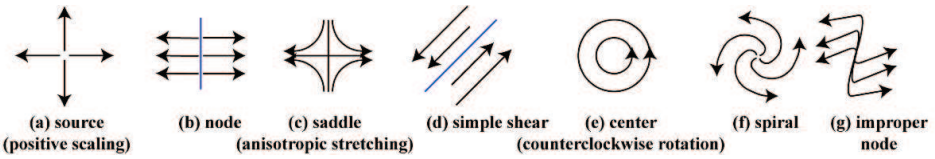


FIGURE 5. Representative flows corresponding to special scenarios on the eigenvalue manifold.

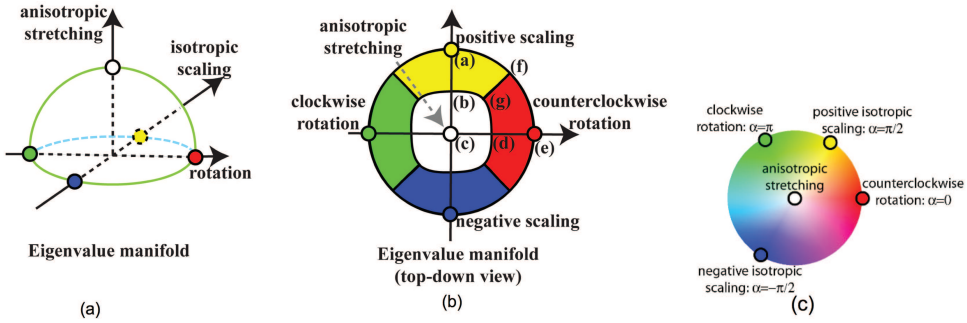


FIGURE 6. Schematic of the five extremal situations in the Eigenvalue manifold of a tensor.

In general, there are five special points in the Eigenvalue manifold that represent the extremal situations: (1) positive scaling ($\gamma_d = 1$, $\gamma_r = \gamma_s = 0$), (2) negative scaling ($\gamma_d = -1$, $\gamma_r = \gamma_s = 0$), (3) counterclockwise rotation ($\gamma_r = 1$, $\gamma_d = \gamma_s = 0$), (4) clockwise rotation ($\gamma_r = -1$, $\gamma_d = \gamma_s = 0$), and (5) anisotropic stretching ($\gamma_s = 1$, $\gamma_d = \gamma_r = 0$). These are shown in the Figure 5.

Figure 6a illustrates the eigenvalue manifold along with the aforementioned special configurations. These configurations lead to a partition of the Eigenvalue manifold based on the spherical geodesic distance. Given two unit vectors v_1 and v_2, the spherical geodesic distance between them is the dot product $1 - v_1 \cdot v_2$. The partition of the Eigenvalue manifold in turn leads to segmentation of the domain into five types of regions:

(1) Counter-clockwise rotation dominated region: $CCWR = \{(x, y)|\gamma_r > \max(\gamma_s, |\gamma_d|)\}$
(2) Clockwise rotation dominated region: $CWR = \{(x, y)|-\gamma_r > \max(\gamma_s, |\gamma_d|)\}$
(3) Positive isotropic scaling dominated region: $PISR = \{(x, y)|\gamma_d > \max(\gamma_s, |\gamma_r|)\}$
(4) Negative isotropic scaling dominated region: $NISR = \{(x, y)| - \gamma_d > \max(\gamma_s, |\gamma_r|)\}$
(5) Anisotropic stretching dominated region: $ASR = \{(x, y)|\gamma_s > \max(|\gamma_d|, |\gamma_r|)\}$

The resulting diagram is illustrated in Figure 6b, in which the boundaries of these regions are highlighted in magenta. The *feature of a tensor field with respect to eigenvalues* consists of points in the domain whose tensor values map to the boundaries between the Voronoi cells in the eigenvalue manifold.

4. Computation of Local and Global Descriptors

Computation of the global descriptors such as Γ & Γ_p and local descriptors such as λ_2 and $\lambda_{2,p}$ is fairly straightforward for both experimental and computational data sets. Once the velocity and pressure gradient fields are obtained these descriptors are extracted at each data-points and can be applied to multiple frames to evaluate the temporal evolution. In this study, data analysis is shown for only one particular time instant.

Evaluation of the vector field topology (ECG) and the tensor field feature needs description. The data sets provided by the experiments or simulation at particular grid nodes are first triangulated. The feature extraction domain is a triangular mesh in either a planar domain or a curved surface. The vector or tensor field is defined at the vertices only. To obtain values at a point on the edge or inside a triangle, a piecewise interpolation scheme is used. For planar domains, this is the well known piecewise linear interpolation scheme [30]. On surfaces, the scheme of Zhang et al. [24, 32] that ensures vector and tensor field continuity in spite of the discontinuity in the surface normal is used.

4.1. Extracting ECG (or vector field topology). Vector field topology for two-dimensional flows consists of fixed points, periodic orbits, and separatrices. An ECG is used to represent vector field topology [16]. To construct an ECG for a vector field represented on a triangular mesh, fixed points such as sources, sink, and saddles are first located and classified based on linearization inside each triangle. Next, periodic orbits are extracted by identifying regions of recurrence in the flow. In the third step, separatrices are computed by tracing streamlines from the saddles in their respective incoming and outgoing directions. This provides edges in the ECG that connect saddles to sources, sinks, and periodic orbits. Finally, edges in the ECG that directly connect between sources, sinks, and periodic orbits are determined by following the forward and reverse directions near periodic orbits that have not been reached by any separatrices. See Chen et al. [16] for details.

4.2. Extracting tensor field feature (Eigenvalue Manifold (EM)). Tensor field feature is computed according to the algorithm of Zhang et al. [34].

Given a tensor field T, the following computation is first performed for every vertex.

(1) Repamaterization, in which γ_d, γ_r, γ_s, and θ are computed.
(2) Normalization, in which γ_d, γ_r, and γ_s are scaled to ensure $\gamma_d^2+\gamma_r^2+\gamma_d^s = 1$.
(3) Eigen-analysis, in which the eigenvalues and eigenvectors are computed.

Next, the feature of the tensor field with respect to the eigenvalues is extracted. This is done by visiting every edge in the mesh to locate possible intersection points with the boundary curves of the Voronoi cells shown in Figure 6. The intersection points are then connected whenever appropriate.

The current implementation of the tensor field feature extraction is on a single processor and is memory and computation intensive. However, parallelization of the approach is possible and straightforward.

5. Results

The results of the flow detection schemes described previously were evaluated for the two data sets: (i) experimentally obtained two dimensional velocity field around a thin wing at a fixed angle of attack, and (ii) data in the symmetry plane from direct numerical simulation of the velocity and pressure fields for flow around a square cylinder. Both vector and tensor field topologies are extracted from the data set. The experimental data is limited to velocity field, and comparison of the vector (Γ and ECG) and tensor field topologies (λ_2 and eigenvalue manifold) are performed. In addition, the original data set is analyzed using low-pass and high-pass filtering to extract features at different scales. Accordingly, any flow variable f can be written as $f = \tilde{f} + f'$, where $\tilde{f}$ and f' represent the low-pass and high-pass filtered data, respectively. A Gaussian filtering operation is used to obtain the low-pass data:

$$\tilde{f}(x,y) = \int_A (f \cdot G)\, dA;\ G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{10}$$

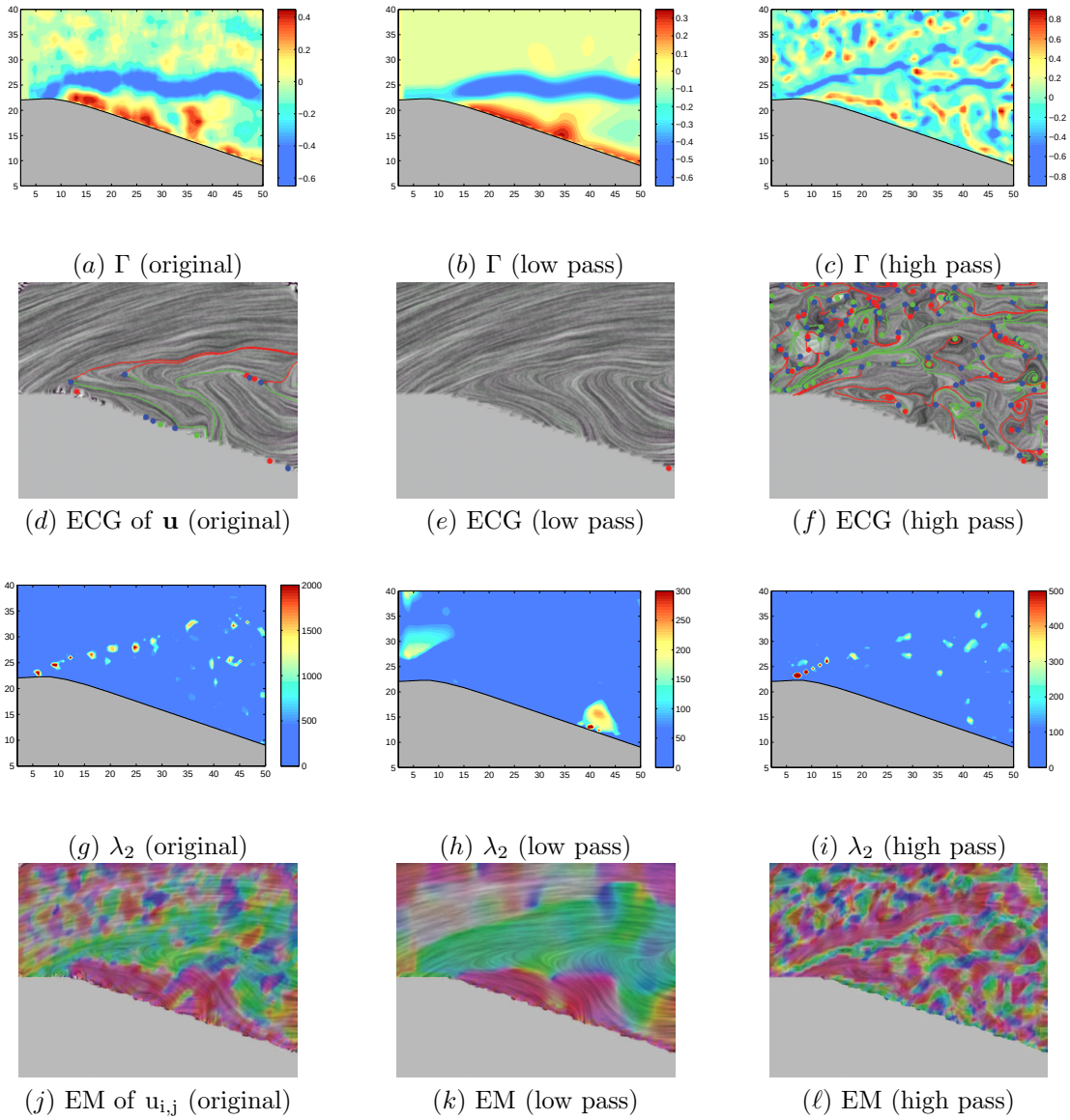
where σ represents the filter width. In this work, the filtered width used is *four* times the local grid resolution. The DNS data set for flow over a square cylinder was evaluated using velocity-based and pressure-gradient based feature extraction techniques. The original data was filtered using the same filtering operation and only the low-pass filtered data is analyzed and presented here. This data set is used to compare the vector and tensor-field topologies based on the velocity and pressure-gradient fields.

5.1. Airfoil Data. The wing flow field analysis is shown in Figure 7, where each row shows the flow features obtained from Γ, ECG, λ_2, and tensor-field eigenvalue topology, respectively. Each column represents the same feature extraction techniques applied to the original, low-pass and high-pass filtered data sets.

Γ Function: The Γ function results (Figure 7a-c) illustrate the detection of well defined swirl that are separated into two main regions, the upper region is a clockwise (negative values) rotating stream that begins at the leading edge of the wing. This represents a flow instability that is generated by this localized separation which is then convected downstream. Below this region, very near the wing surface is a companion region of counterclockwise (positive values) rotating flow. Taken together these regions form a stream of clearly identified counter rotating vortices. The low pass filtered data (Figure 7b) shows apparent smoothing and the stream boundaries are well defined. The high pass filtered data (Figure 7c), on the other hand, shows discrete swirling flow regions of small spatial extent with a much more irregular pattern. Also, remnants of the two main counter rotating streams can be

determined amongst the more randomly positioned swirling vortical flow elements. This shows that the filtered data analysis is capable of detecting smaller scale swirl contained within the larger flow structures.

Note that the Γ-function is obtained by performing spatial integration of the flow quantities (equation 1) at each grid location and represents a *global* detector. However, by separating the length scales over which the velocity field changes (high and low-pass filter), the small scale and large-scale vortical features can be captured. By varying the filter width, a correlation between the flow feature and the filter-width can be obtained to identify the multiscale nature of the turbulent fluid flow.



(a) Γ (original) (b) Γ (low pass) (c) Γ (high pass)

(d) ECG of **u** (original) (e) ECG (low pass) (f) ECG (high pass)

(g) λ_2 (original) (h) λ_2 (low pass) (i) λ_2 (high pass)

(j) EM of $u_{i,j}$ (original) (k) EM (low pass) (ℓ) EM (high pass)

FIGURE 7. A comparison of various techniques for feature extraction applied to the experimental data set of flow over an airfoil with 20^0 angle of attack.

ECG: Figures 7d–f show the vector field topology (ECG) for the original and filtered data sets. The ECGs are shown on top of the flow textures [40] derived from the data sets. For the ECGs shown, sources are represented as green dots, sinks red dots, and saddles blue dots. The separatrices are the colored curves connecting the fixed points (green line connects a saddle to a source, and red line connects the saddle to a sink). For the time-instant shown, no periodic orbits are visible. The connectivity between the fixed points indicate the spatial extent of the various flow events and their interactions. For example, in Figure 7d the two saddles (blue) near the leading edge show connection with the sources (green) on the downstream surface of the wing. This indicates how the flow patterns stretch and evolve through the flow and how surface effects are correlated to flow events far from the surface.

For this flow data, the low-pass filtered ECG does not show any sources, saddles or periodic orbits. One sink is detected in the downstream region. On the other hand, the high-pass filtered data shows a number of fixed points inside and outside the separated flow region. The separatrices show the link between the fixed points. It is observed that inside the separated flow region, the extent of the separatrices is large, indicating that the fixed points detected are correlated with distant flow events. However, outside the separated region (in the free-stream), the flow features are closely correlated by more local events.

This vector-field based feature extraction technique shows similar vortical features as the Γ function. This connectivity information is crucial for multiscale energy cascade mechanisms observed in many turbulent flows. By investigating the statistical nature of fixed points and their correlations, the path associated with energy transfer from large-scale to small scale flow structures can be identified.

λ_2 Method: Figures 7g–i show the tensor field feature as detected by the λ_2 method. As described in the previous section, the λ_2 method is associated with gradient of the vector-field (velocity in this case) and identifies local effects. Consequently, the original data set indicates clockwise rotating flow away from the airfoil surface, similar to the Γ-function. However, it is clear that the Γ-function is able to display well defined swirl flow pattern. While the λ_2 method does detect the strong clockwise rotating flow stream, it only weakly detects the counterclockwise rotation near the surface. Furthermore, the high pass filtered data set does indicate the clockwise stream as a series of small vortical flow elements, in contrast to the high pass data analysis using the Γ function which shows this stream as a nearly continuous region (blue streak) in Figure 7c. These small vortical flow elements seem to be similar to the fixed points identified by the high-pass filtered ECG technique.

Eigenvalue Manifold (EM):. Figures 7j–ℓ show the eigenvalue feature for the original, low-pass and high-pass filtered data sets. Results show strong clockwise rotation (green) along the same region as that detected in the Γ and λ_2 methods. In addition, the near surface region shows a strong counterclockwise rotation (red) coupled with positive isotropic scaling (yellow). This is consistent with the previous methods with essentially the same spatial distribution. It is apparent that the flow over the airfoil is dominated by rotation. Note that the flow is three-dimensional, although the data analyzed is 2D in the $x - y$ plane. Accordingly, as shown earlier, the positive and negative scaling represent mass-flux through the $x - y$ plane. The blue and yellow regions of the tensor field feature thus indicate that strong flux in the z-direction. It can be seen that the clockwise and counterclockwise

rotation regions are separated by positive or negative scaling, representing local three-dimensional effects.

The low pass data sets show that the main swirling streams are well detected providing information on the larger flow structures. The high pass filtered data, as in the case of the λ_2 method, now indicate the small scale flow features with only a slight trace of the larger events. In general, the high pass filtering process allows the detection of the smaller and usually weaker flow events and provides, in this case, a means of examining the extent of low energy background flow events.

5.2. Square Cylinder Data. Figures 8a–h indicate the vector and tensor field feature techniques applied to the planar data for flow over a square cylinder obtained from the direct numerical simulations. Note that the actual computations are full three-dimensional, however, only the two-dimensional data in the symmetry plane is analyzed.

For the square cylinder, only the low-pass filtered data is shown. The goal of this analysis is to compare the various techniques when applied to the velocity-based and pressure-gradient based data sets.

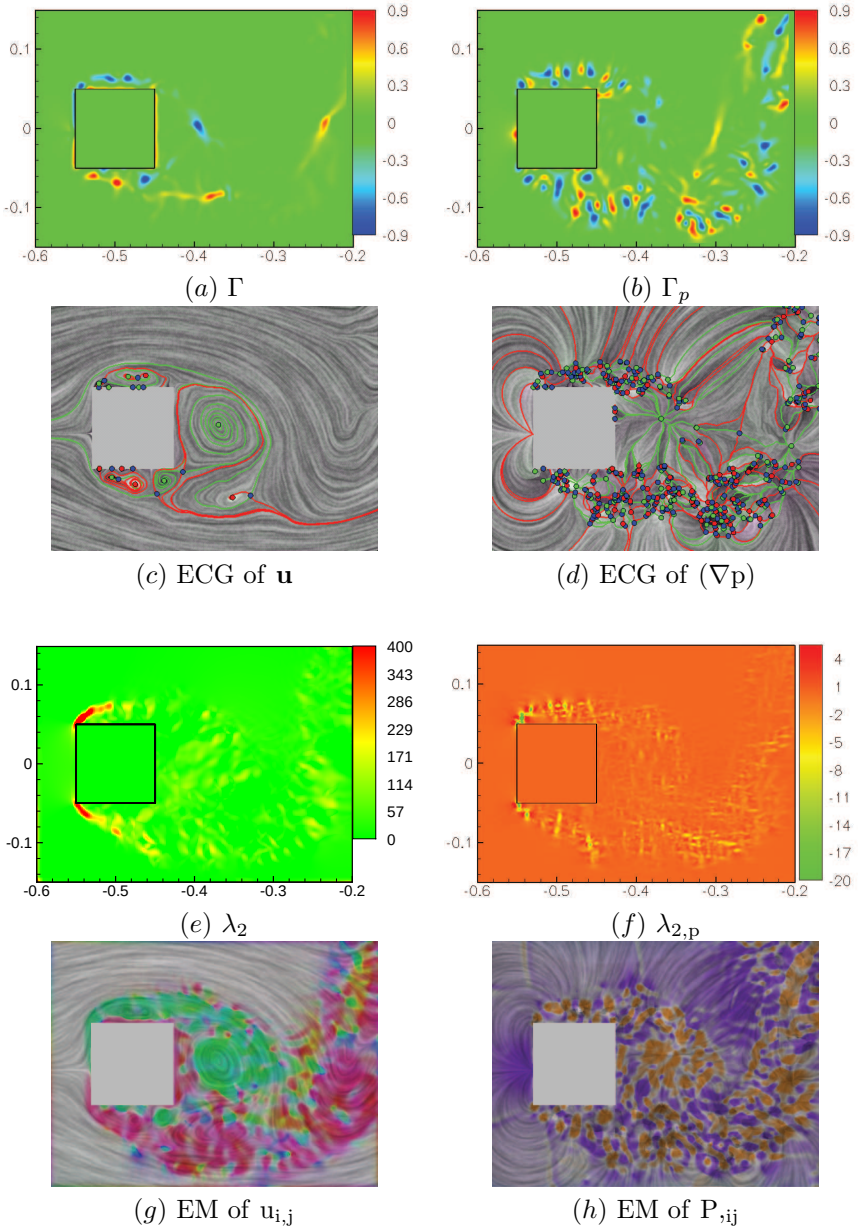
Γ and Γ_p: Figures 8a–b compare the Γ and Γ_p contours for the square cylinder. Both techniques identify the flow separation and swirling regions clearly. The flow separates at the corners of the leading edge. The top corner creates clockwise rotation whereas the bottom-one shows counter-clockwise rotation. The separated flow evolves over the cube surface and a strong wake region is visible downstream of the cylinder. Both techniques identify a strong, clockwise rotation in the wake of the cylinder. It is apparent that the pressure-gradient based Γ_p identifies more features than the velocity based Γ contours. This may be attributed to the fact that Γ_p is based on $(-\nabla P)$, and thus can capture the variations in flow velocity on a smaller scale (local grid size) compared to the velocity vector-based topology.

ECG: Figure 8c–d compares the vector field topology as obtained from the velocity and pressure-gradient fields, respectively. Again, similar swirling patterns as observed by the Γ and Γ_p contours are visible. In this data set sources (green), sinks (red), saddles (blue) and periodic orbits (attracting are red circles) are clearly visible. In addition, the separatrices connecting the fixed points are also shown. Again, the pressure-gradient based topology identifies more fixed points than the velocity field. However, the main vortical structures are identified by both. For example, the large circulation in the wake region (just behind the cylinder) is identified by green dots (source). For the velocity-based ECG, the separatrices show circular paths spiraling around the source. The pressure-gradient based ECG, however, shows lines emanating from the source.

This can be explained by considering a simple case of Rankine vortex (a combination of forced and free vortices):

$$v_\theta = \left\{ \omega r,\ r \le a_c;\ \frac{\omega a_c^2}{r},\ r > a_c \right\},$$

where a_c is the radius of the core of the vortex, ω is the angular rotation associated with the vortex, r is the radial direction, and v_θ is the tangential velocity. The pressure gradient field inside the vortex core is simply given as $\partial p/\partial r = \rho\omega^2 r$. The pressure thus increases with increase in r and the gradient is truly radial. Thus, in a pure vortical flow, the pressure gradient lines are perpendicular to the velocity vector. The separatrices obtained from the pressure-gradient based ECG are seen to be approximately perpendicular to those obtained from the velocity-based ECG



(*a*) Γ (*b*) Γ_p

(*c*) ECG of $\mathbf{u}$ (*d*) ECG of $(\nabla \mathrm{p})$

(*e*) λ_2 (*f*) $\lambda_{2,\mathrm{p}}$

(*g*) EM of $\mathrm{u_{i,j}}$ (*h*) EM of $\mathrm{P_{,ij}}$

FIGURE 8. Comparison of the velocity-based and pressure-gradient based flow descriptors for flow over a square cylinder simulated using direct numerical simulation. Shown is the symmetry plane with a close-up view of the flow field near the square.

(see for example in the wake region). This is indicative of a strong vortical region. The extent of the separatrices roughly scales with the size of the vortical structure.

λ_2 and $\lambda_{2,p}$: Figures 8e–f show the λ_2 contours obtained based on the velocity-gradient and pressure-Hessian tensors. As mentioned earlier, these tensor field topologies identify *locally*, strong regions of swirl. Accordingly, strong swirl regions are obtained near the leading edge corners and also in the wake regions. The pressure Hessian based $\lambda_{2,p}$ shows similar regions of swirl. The contours are more spotty owing to the fact that the topology obtained is based on second derivatives of the pressure.

As discussed in the previous sections, the λ_2 contours try to find pressure minima based on the velocity-gradient based tensor. The temporal, convective, and viscous effects are assumed small to approximately locate the pressure minima. In the $\lambda_{2,p}$ approach, all effects are retained and consequently can locate the vortical regions more accurately.

Eigenvalue Manifold (EM):. Finally, Figures 8g–h show the tensor field topologies obtained from the velocity gradient and pressure Hessian tensors, respectively. The velocity-gradient eigenvalue manifold shows clockwise rotation (green) and counterclockwise rotation (red) regions on the top and bottom surfaces of the cylinder. The wake of the cylinder is dominated by rotation. At the center of the leading edge of the cylinder is a stagnation point representing strong deceleration of the flow, represented by anisotropic stretching. Similarly, outside region of the separated flow is also dominated by anisotropic stretching. The tensor feature shows small regions of positive and negative scaling (blue and yellow regions) and indicate that the flow is mostly two-dimensional.

The eigenvalue manifold of pressure Hessian is shown in Figure 8h. Note that pressure Hessian ($P_{,ij}$) is analyzed, whereas, ($-P_{,ij}$) appears in the equation 6. Accordingly, regions of positive scalings (yellow) in the pressure Hessian correspond to low-pressure regions. This is because in the vicinity of a local minimum of the pressure, the pressure gradient is pointing away from the minimum thus making the minimum a source in the pressure gradient. The flow field, however, is driven by $-\nabla P$ and is towards the vortex center. Similarly, a local maximum in the pressure corresponds to a sink in the pressure gradient (the stagnation point on the leading edge), which resides inside regions of negative scalings (blue).

6. Conclusion

In this work, various techniques, based on vector and tensor fields, to identify multiscale features in turbulent, separated flows were analyzed in detail. These techniques are classified into global and local flow descriptors. The global descriptors are based on spatial integration of flow parameters and thus extract large-scale features. The local techniques are based on the spatial derivatives of flow parameters and identify flow features on the scale of the grid size used to define the flow field. These flow feature extraction techniques were applied to two data sets: (i) experimental velocity field data of flow over a thin airfoil at 20^0 angle of attack, and (ii) direct numerical simulation based data of velocity and pressure-gradient fields for flow over a square cylinder. Both data sets were obtained at flow Reynolds number on the order of 10^4 based on the characteristic size of the bluff body. At these Reynolds numbers, the flow separates and large vortical structures are obtained that convect downstream. The goal of this work was to detect these structures at different scales and compare various techniques.

Two different flow parameters were analyzed. The velocity and pressure-gradient fields were used to obtain the vector field topologies. Two techniques called the

Γ function and the Entity Connection Graph (ECG) were used to deduce the vector field topology. The Γ function maps the degree of rotation rate (or pressure-gradients) to identify local swirl regions, and the ECG combines the Conley theory and Morse decomposition to identify vector field topology consisting of fixed points, periodic orbits, and separatrices connecting them. For both data sets the two techniques detected similar flow features. The Γ function was able to provide the *strength* associated with the vortical structure. The ECG identified recurrent flow features (i.e. fixed points and periodic orbits) in the flow and the separatrices showed the links between these features. The extent of a separatrix connecting two features was found to be roughly proportional to the *scale* of the vortex. From the numerical simulations, the pressure-gradient based topology was obtained and indicated more flow features compared to the velocity-based analysis. The connectivity information between fixed points or vortex centers as provided by the separatrices is an important feature that can be further used to analyze the multiscale energy cascade mechanisms observed in many turbulent flows.

For tensor-field feature the velocity-gradient and pressure Hessian were analyzed. The λ_2 and eigenvalue manifold based techniques were applied to identify the swirling regions. It was observed that the tensor-field feature was capturing vortical structures on the small scale, whereas the extent of the vortices and large-scale features were observed in the vector field topology (Γ and ECG). The eigenvalue manifold decomposed the tensor field into rotation, isotropic scaling, and anisotropic stretching regions indicative of the local flow characteristics. By identifying these regions, it is possible to better understand the dynamics of the separated flows.

Future work will investigate the temporal evolution of the flow features in detail.

Acknowledgments

This work represents a collaborative effort at OSU between the experimental and computational analysis of fluid flow in Mechanical Engineering and scientific visualization in Computer Science. The work is partially supported by NSF, AFOSR, and ONR. James Liburdy and Daniel Morse thank AFOSR for partial support under the grant FA-9550-05-0041. Sourabh Apte and Stephen Snider thank ONR and Dr. Ki-Han Kim for partial support under the grant N000140610697. Eugene Zhang, Guoning Chen, and Zhongzang Lin thank NSF for the partial support under the NSF CAREER grant CCF-0546881. Computations were performed on the San Diego Supercomputing Center's Datastar machine.

References

[1] Y. Hoarau, M. Braza, Y. Ventikos, D. Faghani, and G. Tzabiras, Organized modes and the three dimensional transition to turbulence in the incompressible flow around a NACA0012 wing, J. Fluid Mech., vol. 496, pp. 63–72, 2003.

[2] H. Nishimura and Y. Taniike, Aerodynamic characteristics of fluctuating forces on a circular cylinder, J. Wind Eng., Ind. Aerodynamics, vol. 89, pp. 713–723, 2001.

[3] C. Sicot, S. Auburn, S. Loyer, and P. Devinant, Unsteady characteristics of the static stall of an airfoil subjected to freestream turbulence level up to 16%, Exp. in Fluids, vol. 41, pp. 641–648, 2006.

[4] J. Jeong and F. Hussain, On the identification of a vortex, J. Fluid Mech., vol. 285, pp. 69–94, 1995.

[5] S. Burgmann and C. Brucker and W. Schroder, Scanning PIV measurements of a laminar separation bubble, Exp. Fluids, vol. 41, pp. 319–326, 2006.

[6] C. Weiland and P. Vlachos, Analysis of the Parallel Blade Vortex Interaction with Leading Edge Blowing Flow Control Using the Proper Orthogonal Decompositions, Proceedings FEDSM2007, Joint ASME/JSME Fluids Engineering Conf. July, San Diego, 2007.

[7] B. Dano and J. Liburdy, Vortical structures of a 45^0 inclined pulsed jet in crossflow, AIAA 2006-3542, Fluid Dynamics Conf., San Francisco, CA, 2006.
[8] R. Adrian and K. Christiansen and Z. Liu, Analysis and interpretation of instantaneous velocity fields, Exp. Fluids, vol. 41, pp. 319–326, 2000.
[9] L. Graftieaux and M. Michard and N. Grosjean, Combining PIV, POD and vortex identification algorithms for the study of unsteady turbulent swirling flows, Meas. Sci. Tech., vol. 12, pp. 1422–1429, 2001.
[10] R. S. Laramee, H. Hauser, H. Doleisch, F. H. Post, B. Vrolijk, and D. Weiskopf, The State of the Art in Flow Visualization: Dense and Texture-Based Techniques, Computer Graphics Forum, vol. 23, no. 2, pp. 203–221, 2004.
[11] R. S. Laramee, H. Hauser, L. Zhao, F. H. Post, Topology Based Flow Visualization: The State of the Art, The Topology-Based Methods in Visualization Workshop (TopoInVis 2005), 2006.
[12] J. L. Helman and L. Hesselink, Visualizing Vector Field Topology in Fluid Flows, IEEE Computer Graphics and Applications, vol. 11, no. 3, pp. 36–46, 1991.
[13] G. Scheuermann and H. Krüger and M. Menzel and A. P. Rockwood, Visualizing Nonlinear Vector Field Topology, IEEE Transactions on Visualization and Computer Graphics, vol. 4, no. 2, pp. 109–116, 1998.
[14] X. Tricoche and G. Scheuermann, "Continuous Topology Simplification of Planar Vector Fields," Proceedings IEEE Visualization 2001, pp. 159–166, 2001.
[15] Konrad Polthier and Eike Preuß, 2003, Identifying Vector Fields Singularities using a Discrete Hodge Decomposition, *Mathematical Visualization III, Ed: H.C. Hege, K. Polthier*, Springer Verlag, pp 112-134.
[16] G. Chen, K. Mischaikow, R. S. Laramee, P. Pilarczyk, and E. Zhang, Vector Field Editing and Periodic Orbit Extraction Using Morse Decomposition, IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 4, pp. 769–785, 2007.
[17] A. Globus and C. Levit and T. Lasinski, A Tool for Visualizing the Topology of Three-Dimensional Vector Fields, Proceedings IEEE Visualization '91, pp. 33–40, 1991.
[18] H. Theisel, T. Weinkauf, H.-C. Hege, and H.-P. Seidel, Saddle Connectors–An Approach to Visualizing the Topological Skeleton of Complex 3D Vector Fields, Proceedings IEEE Visualization 2003, pp. 225–232, 2003.
[19] K. Mahrous, J. C. Bennett, G. Scheuermann, B. Hamann, and K. I. Joy, "Topological segmentation in three-dimensional vector fields," IEEE Transactions on Visualization and Computer Graphics, vol. 10(2), pp. 198–205, 2004.
[20] X. Tricoche and T. Wischgoll and G. Scheuermann and H. Hagen, Topology Tracking for the Visualization of Time-Dependent Two-Dimensional Flows, Computers & Graphics, vol. 26, no. 2, pp. 249–257, 2002.
[21] H. Theisel and H.-P. Seidel, "Feature Flow Fields," in Proceedings of the Joint Eurographics - IEEE TCVG Symposium on Visualization (VisSym 03), pp. 141–148, 2003.
[22] H. Theisel, T. Weinkauf, H.-C. Hege, and H.-P. Seidel, "Stream Line and Path Line Oriented Topology for 2D Time-Dependent Vector Fields," Proceedings IEEE Visualization 2004, pp. 321–328, 2004.
[23] C. Garth, X. Tricoche, and G. Scheuermann, Tracking of Vector Field Singularities in Unstructured 3D Time-Dependent Datasets, Proceedings IEEE Visualization 2004, pp. 329–335, 2004.
[24] E. Zhang, K. Mischaikow, and G. Turk, Vector Field Design on Surfaces, ACM Transactions on Graphics, vol. 25, no. 4, pp. 1294–1326, 2006.
[25] Y. Tong, S. Lombeyda, A. Hirani, and M. Desbrun, Discrete Multiscale Vector Field Decomposition, ACM Transactions on Graphics (SIGGRAPH 2003), vol. 22, no. 3, pp. 445–452, 2003.
[26] T. Delmarcelle and L. Hesselink, The Topology of Symmetric, Second-Order Tensor Fields, Proceedings IEEE Visualization '94, 1994.
[27] L. Hesselink, Y. Levy and Y. Lavin, The Topology of Symmetric, Second-Order 3D Tensor Fields, IEEE Transactions on Visualization and Computer Graphics, vol. 3, no. 1, pp. 1–11, 1997.
[28] X. Zheng and A. Pang, Topological Lines in 3D Tensor Fields, Proceedings IEEE Visualization '04, pp. 313–320, 2004.
[29] X. Zheng and B. Parlett and A. Pang, Topological Structures of 3D Tensor Fields, Proceedings IEEE Visualization 2005, pp. 551–558, 2005.

[30] X. Tricoche, G. Scheuermann, and H. Hagen, Tensor Topology Tracking: a Visualization Method for Time-Dependent 2D Symmetric Tensor Fields," in Computer Graphics Forum 20(3) (Eurographics 2001), Sept. 2001, pp. 461–470.
[31] X. Tricoche, G. Scheuermann, and H. Hagen, Topology Simplification of Symmetric, Second-Order 2D Tensor Fields, Hierarchical and Geometrical Methods, Springer, 2003.
[32] E. Zhang, J. Hays, and G. Turk, Interactive Tensor Field Design and Visualization on Surfaces, IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 1, pp. 94–107, 2007.
[33] X. Zheng and A. Pang, 2D Asymmetric Tensor Analysis, IEEE Proceedings on Visualization, pp. 3–10, Oct 2005.
[34] E. Zhang, H. Yeh, Z. Lin, and R. S. Laramee, Asymmetric Tensor Analysis for Flow Visualization, IEEE Transactions on Visualization and Computer Graphics, *to appear*, http://web.engr.orst.edu/∼zhange/images/asymtend.pdf.
[35] K. Mahesh, G. Constantinescu, and P. Moin, A new time-accurate finite-volume fractional-step algorithm for prediction of turbulent flows on unstructured hybrid meshes, J. Comp. Phy., vol. 197, pp. 215–240, 2004.
[36] K. Mahesh, G. Constantinescu, S. Apte, G. Iaccarino, F. Ham, and P. Moin, Large-eddy simulation of reacting turbulent flows in complex geometries, ASME J. App. Mech., vol. 438, pp. 101–128, 2006.
[37] P. Moin and S. Apte, Large eddy simulation of multiphase reacting flows in complex combustors, AIAA J. (special issue on 'Combustion Modeling and LES: Development and Validation Needs for Gas Turbine Combustors), vol. 44, pp. 698–710, 2006.
[38] Apte, S. V., Mahesh, K., Moin, P., & Oefelein, J.C., 2003a, Large-eddy simulation of swirling particle-laden flows in a coaxial-jet combustor. *Int. J. Mult. Flow* **29**, 1311-1331.
[39] Apte, S.V., Mahesh, K., & Lundgren T., 2007 Accounting for Finite-Size Effects in Simulations of Two-Phase Flows *Int. J. Multiphase Flow* accepted for publication.
[40] J. J. van Wijk, Image based flow visualization, in *ACM Transactions on Graphics (SIGGRAPH 02)*, vol. 21, no. 3, Jul 2002, pp. 745–754.

School of Electrical Engineering and Computer Science Oregon State University Corvallis, OR 97331
E-mail: chengu@eecs.oregonstate.edu
URL: http://oregonstate.edu/∼cheng/

School of Electrical Engineering and Computer Science Oregon State University Corvallis, OR 97331
E-mail: lin@eecs.oregonstate.edu

School of Mechanical, Industrial, & Manufacturing Engineering Oregon State University Corvallis, OR 97331
E-mail: morseda@onid.oregonstate.edu

School of Mechanical, Industrial, & Manufacturing Engineering Oregon State University Corvallis, OR 97331
E-mail: snideerst@onid.oregonstate.edu

School of Mechanical, Industrial, & Manufacturing Engineering Oregon State University Corvallis, OR 97331
E-mail: sva@engr.orst.edu
URL: http://me.oregonstate.edu/people/faculty/therm_fluid/apte.html

School of Mechanical, Industrial, & Manufacturing Engineering Oregon State University Corvallis, OR 97331
E-mail: james.liburdy@oregonstate.edu
URL: http://web.engr.oregonstate.edu/∼liburdy/

School of Electrical Engineering and Computer Science Oregon State University Corvallis, OR 97331
E-mail: zhange@eecs.oregonstate.edu
URL: http://web.engr.orst.edu/∼zhange

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 36–46

BROWNIAN MOTION AND ENTROPY GROWTH ON IRREGULAR SURFACES

C. CHEVALIER AND F. DEBBASCH

Abstract. Many situations of physical and biological interest involve diffusions on manifolds. It is usually assumed that irregularities in the geometry of these manifolds do not influence diffusions. The validity of this assumption is put to the test by studying Brownian motions on nearly flat 2D surfaces. It is found by perturbative calculations that irregularities in the geometry have a cumulative and drastic influence on diffusions, and that this influence typically grows exponentially with time. The corresponding characteristic times are computed and discussed. Conditional entropies and their growth rates are considered too.

Key Words. Brownian motion, stochastic processes on manifolds, lateral diffusions.

1. Introduction

Stochastic process theory is one of the most popular tools used in modelling time-asymmetric phenomena, with applications as diverse as economics ([21, 22]), traffic management ([20, 15]), biology ([16, 2, 10, 8]), physics ([23]) and cosmology ([5]). Many diffusions of biological interest, for example the lateral diffusions ([4, 17]), can be modelled by stochastic processes defined on differential manifolds ([12, 13, 9, 18]). In practice, the geometry of the manifold is never known with infinite precision, and it is common to ascribe to the manifold an approximate, mean geometry and to assume irregularities in the geometry have, in the mean, a negligible influence on diffusion phenomena ([4, 1, 3, 6, 19]). The aim of this article is to investigate if this last assumption is indeed warranted.

To this end, we fix a base manifold $\mathcal{M}$ and focus on Brownian motion. We introduce two metrics on $\mathcal{M}$. The first one, g, represents the real, irregular geometry of the manifold; what an observer would consider as the approximate, mean geometry is represented by another metric, which we call $\bar{g}$; to keep the discussion as general as possible, both metrics are allowed to depend on time.

We compare the Brownian motions in the approximate metric $\bar{g}$ to those in the real, irregular metric g by comparing their respective densities with respect to a reference volume measure, conveniently chosen as the volume measure associated to $\bar{g}$. Explicit computations are presented for diffusions on nearly flat 2D surfaces whose geometry fluctuates on spatial scales much smaller than the scales on which these diffusions are observed. We investigate in particular if the densities generated by Brownian motions in the real, irregular metric g coincide on large scales with the densities generated by Brownian motions in the approximate metric $\bar{g}$. We perform a perturbative calculation and find that, generically, these densities differ, even on large scales, and that the relative differences of their spatial

Received by the editors and, in revised form, .
2000 *Mathematics Subject Classification.* 60J65,8J65,60-xx:60Gxx,82Cxx:82C70.

Fourier components grow exponentially in time; on a given surface, the characteristic time τ at which the perturbative terms become comparable (in magnitude) to the zeroth order terms depends on the amplitude ε of the irregularities and on the large scale wave vector k at which diffusions are observed; we find that τ generally scales as $-\left(\nu^{-2}\ln(\varepsilon/\nu^{1/2})\right)\times\left(1/\mid K^*\mid^2\chi\right)$, where χ is the diffusion coefficient and $\nu=\mid k\mid/\mid K^*\mid$, K^* being a typical wave-vector characterizing the metric irregularities. Our general conclusion is that geometry fluctuations have a cumulative effect on Brownian motion and that their influence on diffusions cannot be neglected.

2. Brownian motions on a manifold

2.1. Brownian motion in a time-independent metric. Let $\mathcal{M}$ be a fixed real base manifold of dimension d. Let g be a time-independent metric on $\mathcal{M}$. This metric endows $\mathcal{M}$ with a natural volume measure which will be denoted hereafter by dVol_g. If $\mathcal{C}$ is a chart on $\mathcal{M}$ with coordinates $x=(x^i), i=1,...,d$, integrating against dVol_g comes down to integrating against $\sqrt{\mathrm{det}g_{ij}}d^dx$, where the g_{ij}'s are the components of g in the coordinate basis associated to $\mathcal{C}$.

There is a canonical definition of a Brownian motions on $\mathcal{M}$ equipped with metric g ([14, 9, 11, 18]). Quite intuitively, these Brownian motions are defined through the diffusion equation obeyed by their densities n with respect to dVol_g. Given an arbitrary positive diffusion constant χ, this equation reads:

$$\partial_t n=\chi\Delta_g n, \tag{1}$$

where Δ_g is the Laplace-Beltrami operator associated to g ([7]); given a chart $\mathcal{C}$ with coordinates x, one can write:

$$\Delta_g n=\frac{1}{\sqrt{\mathrm{det}g_{kl}}}\partial_i\left(\sqrt{\mathrm{det}g_{kl}}\,g^{ij}\partial_j n\right), \tag{2}$$

where ∂_i represents partial derivation with respect to x^i and the g^{ij}'s are the components of the inverse of g in the coordinate basis associated to $\mathcal{C}$. Observe that one of the reasons why this definition makes sense is that the diffusion equation (1) conserves the normalization of n with respect to dVol_g.

2.2. Brownian motion in a time-dependent metric. The preceding definition of Brownian motion cannnot be used in this case because the diffusion equation (1) does not conserve the normalization of $n(t)$ with respect to the volume measure $\mathrm{dVol}_{g(t)}$ associated to a time-dependent metric. To proceed, we introduce an arbitrary, time-independent metric γ on $\mathcal{M}$, denote by $\mu_{g(t)|\gamma}$ the density of $\mathrm{dVol}_{g(t)}$ with respect to dVol_γ, and define the Brownian motion in the time-dependent metric $g(t)$ as the stochastic process whose density n with respect to $\mathrm{dVol}_{g(t)}$ obeys the following generalized diffusion equation:

$$\frac{1}{\mu_{g(t)|\gamma}}\partial_t\left(\mu_{g(t)|\gamma}n\right)=\chi\Delta_{g(t)}n. \tag{3}$$

Given an arbitrary coordinate system (x), equation (3) transcribes into:

$$\partial_t\left(\sqrt{\mathrm{det}g_{kl}}n\right)=\chi\partial_i\left(\sqrt{\mathrm{det}g_{kl}}\,g^{ij}\partial_j n\right), \tag{4}$$

which shows that the Brownian motion in $g(t)$ does not actually depend on γ. Moreover,

$$\begin{aligned}\frac{d}{dt}\int_{\mathcal{M}}\mathrm{dVol}_{g(t)}n &= \frac{d}{dt}\int_{\mathcal{M}}\mathrm{dVol}_{\gamma}\mu_{g(t)|\gamma}n\\ &= \int_{\mathcal{M}}\mathrm{dVol}_{\gamma}\partial_t\left(\mu_{g(t)|\gamma}n\right)\\ &= \chi\int_{\mathcal{M}}\mathrm{dVol}_{\gamma}\mu_{g(t)|\gamma}\Delta_{g(t)}n\\ &= \chi\int_{\mathcal{M}}\mathrm{dVol}_{g(t)}\Delta_{g(t)}n\\ &= 0. \qquad (5)\end{aligned}$$

Thus, contrary to (1), equation (3) conserves the normalization of $n(t)$.

2.3. Entropies of Brownian motion in a time-dependent metric. Let n and $\tilde{n}$ be two solutions of (3). We define the time-dependent conditional entropy $S_{n|\tilde{n}}$ of n with respect to $\tilde{n}$ by:

$$S_{n/\tilde{n}}(t) = -\int_{\mathcal{M}}\mathrm{dVol}_{g(t)}n\ln(\frac{n}{\tilde{n}}). \qquad (6)$$

This entropy is a non decreasing function of t. This can be seen by the following calculation. One can write:

$$S_{n|\tilde{n}}(t) = -\int_{\mathcal{M}}\mathrm{dVol}_{\gamma}\ \mu_{g(t)|\gamma}n\ \ln(\frac{\mu_{g(t)|\gamma}n}{\mu_{g(t)|\gamma}\tilde{n}}), \qquad (7)$$

which leads to

$$\begin{aligned}\frac{dS_{n|\tilde{n}}}{dt} &= -\int_{\mathcal{M}}\mathrm{dVol}_{\gamma}\left(\partial_t(\mu_{g(t)|\gamma}n)\ln(\frac{\mu_{g(t)|\gamma}n}{\mu_{g(t)|\gamma}\tilde{n}})\right.\\ &+ \left.\mu_{g(t)|\gamma}n\frac{\mu_{g(t)|\gamma}\tilde{n}\partial_t(\mu_{g(t)|\gamma}n)-\mu_{g(t)|\gamma}n\partial_t(\mu_{g(t)|\gamma}\tilde{n})}{\mu^2_{g(t)|\gamma}n\tilde{n}}\right). \qquad (8)\end{aligned}$$

Equation (3) can then be used to transform all temporal derivatives into spatial ones and one obtains:

$$\frac{dS_{n|\tilde{n}}}{dt} = -\chi\int_{\mathcal{M}}\mathrm{dVol}_{g(t)}\left(\Delta_{g(t)}n\left(\ln(\frac{n}{\tilde{n}})+1\right)-\frac{n}{\tilde{n}}\Delta_{g(t)}\tilde{n}\right). \qquad (9)$$

Integrating by parts delivers:

$$\frac{dS_{n|\tilde{n}}}{dt} = +\chi\int_{\mathcal{M}}\mathrm{dVol}_{g(t)}\tilde{n}\left(\nabla(\ln\frac{n}{\tilde{n}})\right)^2, \qquad (10)$$

which proves the expected result. Conditional entropies of Brownian motions thus obey a very simple H-theorem, even in time-dependent geometries.

The Gibbs entropy $S_G[n]$ of a density n is defined by

$$S_G[n](t) = -\int_{\mathcal{M}}\mathrm{dVol}_{g(t)}n\ln n. \qquad (11)$$

The H-theorem above applies to $S_G[n]$ only if the $\tilde{n}=1$ is a solution of the transport equation (3). This is automatically the case if the metric g is time-independent, but may not be true in time-dependent metrics. Note also that a uniform density is not normalizable on non compact manifolds.

3. How to compare Brownian motions in different metrics

Let $\mathcal{M}$ be a real differential manifold of dimension d. We first introduce on $\mathcal{M}$ a metric $\bar{g}(t)$ which describes what an observer would consider as the approximate, mean geometry of the manifold. The real, irregular geometry of $\mathcal{M}$ is described by a different metric $g(t)$.

Consider an arbitrary point O in $\mathcal{M}$ and let B_t be the Brownian motion in $g(t)$ that starts at O. The density n of B_t with respect to $\mathrm{dVol}_{g(t)}$ obeys the diffusion equation:

$$\frac{1}{\mu_{g(t)|\gamma}}\partial_t\left(\mu_{g(t)|\gamma}n\right)=\chi\Delta_{g(t)}n. \tag{12}$$

We denote by $\bar{B}_t$ the Brownian motion in $\bar{g}(t)$ that starts at point O and by $\bar{n}$ its density with respect to $\mathrm{dVol}_{\bar{g}(t)}$; this density obeys:

$$\frac{1}{\mu_{\bar{g}(t)|\gamma}}\partial_t\left(\mu_{\bar{g}(t)|\gamma}\bar{n}\right)=\chi\Delta_{\bar{g}(t)}\bar{n}. \tag{13}$$

We will compare the two Brownian motions by comparing on large scales their respective densities with respect to a reference volume measure on $\mathcal{M}$. From an observational point of view, the best choice is clearly $\mathrm{dVol}_{\bar{g}(t)}$, the volume measure associated to the approximate, mean geometry of the manifold. The density N of B_t with respect to $\mathrm{dVol}_{\bar{g}(t)}$ is given in terms of n by:

$$N=\mu_{g(t)|\bar{g}(t)}n, \tag{14}$$

where $\mu_{g(t)|\bar{g}(t)}$ is the density of $\mathrm{dVol}_{g(t)}$ with respect to $\mathrm{dVol}_{\bar{g}(t)}$. The transport equation obeyed by N can be deduced from (12) and reads:

$$\frac{1}{\mu_{g(t)|\gamma}}\partial_t\left(\mu_{\bar{g}(t)|\gamma}N\right)=\chi\Delta_{g(t)}\left(\frac{1}{\mu_{g(t)|\bar{g}(t)}}N\right). \tag{15}$$

In a chart $\mathcal{C}$ with coordinates (x), (14) transcribes into:

$$N(t,x)=\frac{\sqrt{\det g_{ij}(t,x)}}{\sqrt{\det\bar{g}_{ij}(t,x)}}n(t,x) \tag{16}$$

and (15) becomes:

$$\partial_t\left(\sqrt{\det\bar{g}_{kl}(t,x)}N(t,x)\right)=\chi\partial_i\left(\sqrt{\det g_{kl}(t,x)}g^{ij}(t,x)\partial_j\frac{\sqrt{\det\bar{g}_{kl}(t,x)}}{\sqrt{\det g_{kl}(t,x)}}N(t,x)\right). \tag{17}$$

Let N and $\tilde{N}$ be the densities with respect to $\mathrm{dVol}_{\bar{g}(t)}$ corresponding to two solutions n and $\tilde{n}$ of equation (3). The conditional entropy of n with respect to $\tilde{n}$ can also be written:

$$S_{n|\tilde{n}}(t)=-\int_{\mathcal{M}}\mathrm{dVol}_{\bar{g}(t)}N\ln(\frac{N}{\tilde{N}}). \tag{18}$$

This entropy can thus be also interpreted as the conditional entropy of N with respect to $\tilde{N}$ on the manifold equiped with metric $\bar{g}(t)$. Note however that the Gibbs entropy of n in $g(t)$ does not coincide with the Gibbs entropy of N in $\bar{g}(t)$.

The main question investigated in this article is: how does the density N obeying (15) differ on large scales from the density $\bar{n}$ obeying (13)? Since this question is extremely difficult to solve in its full generality, we now concentrate on nearly flat 2D surfaces.

4. Brownian motions on nearly flat 2D surfaces

4.1. The problem. We choose $\mathbb{R}^2$ as base manifold $\mathcal{M}$ and retain $\bar{g} = \eta$, the flat Euclidean metric on $\mathbb{R}^2$. The real, irregular metric of the manifold is still denoted by $g(t)$ and we define $h(t)$ by $g^{-1}(t) = \eta^{-1} + \varepsilon h(t)$, where ε is a small parameter (infinitesimal) tracing the nearly flat character of the surface. From now on, we will use the metric η (*resp.* the inverse of η) to lower (*resp.* raise) all indices.

Let us choose a chart $\mathcal{C}$ where $\eta_{ij} = \mathrm{diag}(1,1)$. The tensor field $h(t)$ is then represented by its components $h^{ij}(t,x)$. A particularly simple but very illustrative form for these components is:

$$h^{ij}(t,x) = \sum_{nn'} h^{ij}_{nn'} \cos\left(\omega_{n'} t - k_n.x + \phi_{nn'}\right), \tag{19}$$

where $k_n.x = k_{n\,1}x^1 + k_{n\,2}x^2$ and both integer indices run through arbitrary finite sets. This choice has the double advantage of leading to conclusions which are sufficiently robust to remain qualitatively valid for all sorts of physically interesting perturbations h while making all technical aspects of the forthcoming computations and discussions as simple as possible. The *Ansatz* (19) will therefore be retained in the remainder of this article. Let us remark that perturbations $h(t)$ proportional to η amount to a simple modification of the conformal factor linking the 2D metric $g(t)$ to the flat metric η.

Equation (17) reads, in the chart $\mathcal{C}$:

$$\partial_t N = \chi \partial_i \left(\sqrt{\mathrm{det} g_{kl}(t,x)} g^{ij}(t,x) \partial_j \frac{N}{\sqrt{\mathrm{det} g_{kl}(t,x)}} \right) \tag{20}$$

or, alternately,

$$\partial_t N = \chi \partial_i \left(g^{ij}(t,x) \left(\partial_j N - N \partial_j l \right) \right), \tag{21}$$

where

$$l(t,x) = \ln \sqrt{\mathrm{det} g_{kl}(t,x)}. \tag{22}$$

4.2. General perturbative solution. The solution of (21) will be searched for as a perturbation series in the amplitude ε of the fluctuations:

$$N(t,x) = \sum_{m \in N} \varepsilon^m N_m(t,x). \tag{23}$$

Setting to 0 both coordinates of the point O where the diffusion starts from, we further impose, for all x, that $N_0(0,x) = \delta(x)$ and $N_m(0,x) = 0$ for all $m > 0$.

The function $l(t,x)$ can be expanded in ε, so that $l(t,x) = \sum_{m \in N} \varepsilon^m l_m(t,x)$ and one finds, for the first three contributions:

$$\begin{aligned} l_0(t,x) &= 0 \\ l_1(t,x) &= -\frac{1}{2}\, \eta_{ij} h^{ij}(t,x) \\ l_2(t,x) &= \frac{1}{4} \eta_{ik}\eta_{jl} h^{ij}(t,x) h^{kl}(t,x). \end{aligned} \tag{24}$$

Equation (21) can then be rewritten as the system

$$\partial_t N_m = \chi \Delta_\eta N_m + \chi S_m[h, N_r], m \in \mathbb{N}, r \in \mathbb{N}_{m-1} \tag{25}$$

where the source term S_m is a functional of the fluctuation h and of the contributions to N of order strictly lower than m. In particular,

$$\begin{aligned} S_0 &= 0 \\ S_1 &= \partial_i \left(h^{ij}\partial_j N_0 + \frac{1}{2} N_0 \eta^{ij}\eta_{kl}\partial_j h^{kl} \right) \\ S_2 &= \partial_i \left(h^{ij}\partial_j N_1 + \frac{1}{2}\left(N_0 h^{ij} + N_1 \eta^{ij}\right)\eta_{kl}\partial_j h^{kl} - \frac{1}{4} N_0 \eta^{ij}\eta_{mk}\eta_{nl}\partial_j \left(h^{mn}h^{kl}\right) \right). \end{aligned} \tag{26}$$

Two remarks are now in order. Taken together, $S_0(t,x) = 0$ and $N_0(t,x) = \delta(x)$ imply that N_0 coincides with the Green function of the standard diffusion equation on the flat plane:

$$N_0(t,x) = \frac{1}{4\pi\chi t} \exp\left(-\frac{x^2}{4\chi t}\right). \tag{27}$$

Moreover, the fact that S_m is a divergence for all m implies that the normalizations of all N_m's are conserved in time. The initial condition $N_m(0,x) = 0$ for all x and $m > 0$ then implies that all N_m's with $m > 0$ remain normalized to zero and only contribute to the local density of particles, and not to the total density. This is perfectly coherent with the fact that N_0 is normalized to unity.

Define now spatial Fourier transforms by

$$\hat{f}(t,k) = \int_{\mathbb{R}^2} f(t,x)\exp(-ik.x)d^2x, \tag{28}$$

where $k.x = k_1x^1 + k_2x^2$. A direct calculation then delivers:

$$\hat{S}_1(t,k) = -k_i \int_{\mathbb{R}^2} A^i(t,k,k')\hat{N}_0(t,k-k')d^2k' \tag{29}$$

where

$$\hat{N}_0(t,k) = \exp\left(-\chi k^2 t\right) \tag{30}$$

and

$$A^i(t,k,k') = (k_j - k'_j)\hat{h}^{ij}(t,k') + \frac{1}{2}\,\eta^{ij}k'_j\eta_{kl}\hat{h}^{kl}(t,k'). \tag{31}$$

The first order density fluctuation N_1 is then obtained by solving equation (25) with (29) as source term, taking into account the initial condition $N_1(0,x) = 0$ for all x. One thus obtains:

$$\hat{N}_1(t,k) = I_1(t,k)\exp(-\chi k^2 t) \tag{32}$$

with

$$I_1(t,k) = \int_0^t \hat{S}_1(t',k)\exp(\chi k^2 t')dt'. \tag{33}$$

Equation (26) then gives:

$$\begin{aligned} \hat{S}_2(t,k) &= -k_i \int_{\mathbb{R}^2} A^i(t,k,k')\hat{N}_1(t,k-k')d^2k' \\ &+ \int_{\mathbb{R}^2\times\mathbb{R}^2} B^i(t,k',k'')\hat{N}_0(t,k-k')d^2k'd^2k'' \end{aligned} \tag{34}$$

with

$$B^i(t,k,k') = \frac{1}{2}\, k'_j \eta_{kl} \hat{h}^{kl}(t,k') \hat{h}^{ij}(t,k-k') - \frac{1}{4}\, \eta^{ij} k'_j \eta_{mk} \eta_{nl} \left(\hat{h}^{mn}(t,k') \hat{h}^{kl}(t,k-k') + \hat{h}^{kl}(t,k') \hat{h}^{mn}(t,k-k') \right). \quad (35)$$

The second order density fluctuation N_2 then reads:

$$\hat{N}_2(t,k) = I_2(t,k) \exp(-\chi k^2 t) \quad (36)$$

with

$$I_2(t,k) = \int_0^t \hat{S}_2(t',k) \exp(\chi k^2 t') dt'. \quad (37)$$

4.3. How the irregularities influence diffusions.

4.3.1. First order terms. Let us now insert *Ansatz* (19) in the above expressions (29) and (32) for $\hat{S}_1$ and $\hat{N}_1$. One finds:

$$\hat{S}_1(t,k) = \sum_{nn'\sigma} A^{\sigma}_{nn'}(k) \exp\left(i\sigma(\omega_{n'} t + \phi_{nn'}) - (k+\sigma k_n)^2 \chi t \right) \quad (38)$$

with

$$A^{\sigma}_{nn'}(k) = -\frac{1}{2} \left[k_i (k_j + \sigma k_{nj}) h^{ij}_{nn'} - \frac{1}{2}\, \sigma \eta^{ij} k_i k_{nj} \eta_{kl} h^{kl}_{nn'} \right] \quad (39)$$

and $\sigma \in \{-1,+1\}$. This leads to:

$$\frac{\hat{N}_1(t,k)}{\hat{N}_0(t,k)} = \sum_{\sigma} \left\{ \sum_{(n,n') \notin \Sigma^{\sigma}(k)} I^{\sigma}_{nn'}(k) \left[\exp\left(\sigma i \omega_{n'} t - (k_n^2 + 2\sigma k.k_n)\chi t \right) - 1 \right] + t \sum_{(n,n') \in \Sigma^{\sigma}(k)} A^{\sigma}_{nn'}(k) \exp\left(i\phi_{nn'} \right) \right\}. \quad (40)$$

with

$$I^{\sigma}_{nn'}(k) = \frac{A^{\sigma}_{nn'}(k) \exp(i\phi_{nn'})}{i\sigma\omega_{n'} + (k^2 - (k+\sigma k_n)^2)\, \chi}, \quad (41)$$

and $\Sigma^{\sigma}(k) = \left\{ (n,n'), \sigma i \omega_{n'} + \left(k^2 - (k \sigma k_n)^2 \right) \chi = 0 \right\}$. Note that both sets are disjoint, unless there is an (n,n') for which $k_n = 0$ and $\omega_{n'} = 0$.

This expression characterizes how Brownian motions in the irregular metric differ, at first order, from Brownian motions on the flat Euclidean plane. The dependence on the wave vector k indicates that the influence of the irregularities varies with the spatial scale at which the diffusion is observed. Two opposite situations are particularly worth commenting upon. Take a certain k_n and consider $\hat{N}_1$ at scales characterized by wave vectors much smaller than k_n, say $\mid k \mid = \mid k_n \mid O(\nu)$, where ν is an infinitesimal (small parameter). Neglecting the contributions of the frequencies $\omega_{n'}$, the amplitudes $A^{\pm}_{nn'}(k)$ typically scale as $\mid k \mid\mid k_n \mid$, so that the $I^{\pm}_{nn'}(k)$'s scale as $O(\nu)$. Note however that perturbations h proportional to η do not obey this typical scaling, but rather $A^{\pm}_{nn'}(k) \sim k^2$, and $I^{\pm}_{nn'}(k) \sim O(\nu^2)$. The sets $\sigma^{\pm}$ are empty and the time-dependence of $\mid \hat{N}_1/\hat{N}_0 \mid$ is controled by the real exponentials in (40), which essentially decrease as $\exp(-k_n^2 \chi t)$. The first order relative contribution $\varepsilon \hat{N}_1/\hat{N}_0$ thus tends towards a quantity $L_1(k)$ which is linear in the $I^{\pm}_{nn'}(k)$; the typical relaxation time is $\tau_1 \sim 1/(\chi k_n^2)$, which is much smaller than the diffusion time $1/(\chi k^2)$ associated to scale k. Moreover, the limit $L_1(k)$ scales as

$O(\varepsilon\nu)$, except for perturbations h proportional to η, for which it scales as $O(\varepsilon\nu^2)$; $L_1(k)$ is therefore always much smaller than ε and, in particular, tends to zero with ν *i.e.* as the scale separation tends to infinity. The effect of the k_n Fourier mode on scales characterized by a wave vector k verifying $\mid k \mid \ll \mid k_n \mid$ is thus in practice negligible.

Consider now the opposite case, *i.e.* $\mid k \mid$ comparable to, or larger than $\mid k_n \mid$. Neglecting again the contribution of the frequencies $\omega_{n'}$, the amplitudes $A^{\pm}_{nn'}(k)$ then scale as k^2, and $I^{\pm}_{nn'}(k) \sim \mid k \mid / \mid k_n \mid$. Let now θ be the angle between k and k_n and suppose, to simplify the discussion, that $\cos\theta$ does not vanish. At least one of the exponentials in (40) will then be an increasing function of time provided $\mid k \mid > \mid k_n \mid /(2\cos\theta)$. Take for example $k = k_n$; the second exponential in (40) then increases with a characteristic time-scale $1/(k^2\chi)$. This means that the first order contributions of the irregularities to the density actually become comparable to unity at this scale at characteristic times $\tau_1 \sim -(\ln\varepsilon)/(k^2\chi)$; this time probably also signals the break down of the perturbative expansion in ε for the scale $k = k_n$. As for the linear terms in t appearing on the right-hand side of (40), they actually contribute to $\hat{N}_1/\hat{N}_0$ if at least one of the sets $\sigma^{\pm}(k)$ is not empty. This condition is realized if $\omega_{n'} = 0$ and $\mid k \mid = \mid k_n \mid /(2\cos\theta)$.

The conclusion of this discussion is that, at first order in the amplitude of the perturbation h, a given Fourier mode k_n of h dramatically influences diffusions on scales characterized by wave vectors with modulus comparable to or larger than $\mid k_n \mid$, but has a negligible influence on scales characterized by wave vectors with modulus much smaller than $\mid k_n \mid$. We will now show that this conclusion cannot be extended to all perturbation orders and that taking into account terms of orders higher than 1 proves that h generally influences diffusions on all scales.

4.3.2. Second order terms. It is straightforward to obtain from equations (34), (36) , (40) and (30) explicit expressions for $\hat{S}_2(t,k)$ and $\hat{N}_2(t,k)/\hat{N}_0(t,k)$. These are extremely complicated and do not warrant full reproduction here.

Of interest is that $\hat{N}_2/\hat{N}_0$ contains contributions whose amplitudes potentially grows exponentially in time. One of these reads

$$D_1(t,k) = \sum_{n,n',\sigma_1,\sigma_2,\sigma_3} I^{\sigma_1}_{nn'}(k+\sigma_2 k_p) A^{\sigma_2}_{pp'}(k) J^{\sigma_1\sigma_3}_{nn'pp'}(k) \times$$

$$\left[\exp\left(i(\sigma_1\omega_{n'} + \sigma_2\omega_{p'})t - \left((k_n + \sigma_1\sigma_2 k_p)^2 + 2k.(k_n + \sigma_1\sigma_2 k_p)\right)\chi t\right) - 1\right] \tag{42}$$

with $\sigma_i = \pm 1$ $(i = 1,2,3)$ and

$$J^{\sigma_1\sigma_3}_{nn'pp'}(k) = \frac{\exp\left(i\sigma_1\sigma_3\phi_{pp'}\right)}{i\sigma_1(\omega_{n'} + \sigma_3\omega_{p'}) + (k^2 - (k + \sigma_1(k_n + \sigma_3 k_p))^2)\chi}. \tag{43}$$

The right-hand sides of (42) contains four exponentials of given (n, n', p, p'); these involve the wave vectors $K^{\pm}_{np} = k_n \pm k_p$. Let us for the moment ignore the factors in front of these exponentials. Let k be an arbitrary wave vector and let $\theta^{\pm}$ be the angle between k and $K^{\pm}_{np}$. Each of the conditions $2k \mid \cos\theta^{\pm} \mid > \mid K^{-}_{np} \mid$ makes one of the four exponentials an increasing function of t. At second order, the spatial scales at which diffusions are influenced by the perturbation h are thus determined, not by the k_n's, but by the combinations $K^{\pm}_{np} = k_n \pm k_p$. Indeed, quite generally, the temporal behaviour of terms of order q, $q \geq 1$, will be determined by combinations of q wave vectors k_n. For perturbations h with a rich enough spectrum, these combinations correspond to all sorts of spatial scales and, in particular, to scales much larger than those over which h itself varies. Thus, h will generally influence diffusions on all spatial scales.

Let us elaborate quantitatively on this conclusion by further exploring the behaviour of $D_1(t,k)$. Suppose for example that the moduli of all k_n's are of the same order of magnitude, say k^*, but that there are some n and p for which $\mid K^-_{np} \mid \sim K^* O(\nu)$, where $\nu << 1$. The condition introduced above, which ensures that one of the exponentials involving K^-_{np} grows with t, then translates into $\mid k \mid > (2/\cos\theta^-)K^* O(\nu)$, and is realized for $\mid k \mid = K^* O(\nu)$ provided $\cos\theta^- \lesssim 1$ Let us check now that the factors in front of the exponentials do not tend towards zero with ν. Ignoring as before the influence of the frequencies ω_q, the quantity $I^-_{nn'}(k+k_p)$ (see (41)) scales as $A^-_{nn'}(k+k_p)/k_p^2$ *i.e.* as $k_p^2/k_p^2 = 1$. The quantity $\tilde{J}^-_{nn'pp'}(k)$ scales as $(Q_{np}(k))^{-1} = \left[2k.K^-_{np} - (K^-_{np})^2\right]^{-1}$. The factor in front of the exponential thus scales as $\mid k \mid\mid k_p \mid (Q_{np}(k))^{-1}$ for perturbations h not proportional to η, and as $k^2(Q_{np}(k))^{-1}$ otherwise. Taking into account that $\mid k \mid \sim \mid K_{np} \mid$ and putting $\cos\theta^- = 1$ to simplify the discussion, one finds that the factor in front of the exponentials scales as $\mid k_p \mid / \mid k \mid = O(1/\nu)$ if h is not proportional to η and as $O(1)$ otherwise. This factor therefore does not tend to zero with the separation scale parameter ν. Actually, for perturbations which are not proportional to η, this factor tends to infinity as ν tends to zero, a fact which only increases the influence of h on diffusions.

These estimates can be used to evaluate some characteristic times. For perturbations proportional to η, the second order term $\varepsilon^2 D_1$ reaches unity after a characteristic time $\tau_2^\eta \sim -(2/\nu^2 K^{*2}\chi)\ln\varepsilon$; for perturbations not proportional to η, the corresponding characteristic time is $\tau_2 \sim -(2/\nu^2 K^{*2}\chi)\ln(\varepsilon/\nu^{1/2}) << \tau_2^\eta$. These characteristic times are probably upper bound for the time at which the perturbation expansion ceases to be valid for scale k.

5. Influence of the irregularities on the entropies

The influence of generic metric irregularities on conditional entropies is technically extremely difficult to investigate in detail. We therefore restrict our discussion by considering only time-independent perturbations h proportional to η and write:

$$h^{ij}(x) = \sum_n a_n \eta^{ij} \cos\left(k_n.x - \phi_n\right). \tag{44}$$

Let us focus on the Gibbs entropy $S_G[n]$ of the density n evaluated in Section 4. This entropy reads

$$S_G[n](t) = -\int_{\mathbb{R}^2} d^2x\ N(t,x) \ln\left(\frac{N(t,x)}{\sqrt{\det g(t,x)}}\right). \tag{45}$$

The perturbative expansion of both g and N, together with the normalization conditions $\int_{\mathbb{R}^2} N_m d^2x = 0$ for $m = 1, 2$, leads to:

$$S_G[n](t) = \sum_{m\in N} \varepsilon^m S_{Gm}[n](t) \tag{46}$$

with

$$\begin{aligned}
&S_{G0}[n](t) = -\int_{\mathbb{R}^2} d^2x\ N_0(t,x)\ln N_0(t,x),\\
&S_{G1}[n](t) = -\int_{\mathbb{R}^2} d^2x \left(N_1(t,x)\ln N_0(t,x) - N_0(t,x) l_1(x)\right),\\
&S_{G2}[n](t) =\\
&-\int_{\mathbb{R}^2} d^2x \left(N_2(t,x)\ln N_0(t,x) - N_0(t,x) l_2(x) - N_1(t,x) l_1(x) + \frac{N_1^2(t,x)}{2N_0(t,x)}\right).
\end{aligned} \tag{47}$$

Expression (27) for N_0 leads to $S_{G0}[n](t) = 1 + \ln(4\pi\chi t)$, which is, as expected, an increasing function of t. This function is also strictly positive. A direct computation shows that N_1 is an uneven function of x. SInce N_0 is even in x, so is $\ln N_0$ and the contribution of $N_1 \ln N_0$ to $S_{G1}[n]$ vanishes identically. One finds, using (24), that:

$$S_{G1}[n](t) = -\sum_n a_n \cos\phi_n \exp\left(-k_n^2 \chi t\right). \tag{48}$$

The first order contribution to the Gibbs entropy may thus be a decreasing or an increasing function of time, and its sign is not fixed either. Each term in the sum tends to zero on a characteristic time $T_n = 1/(\chi k_n^2)$. Suppose the diffusion is observed at scale k with $\mid k \mid = \mid k_n \mid O(\nu)$. The relaxation time T_n is then much smaller than the typical diffusive time $T = 1/(\chi k^2)$ at scale k and the first order contribution to $S_G[n]$ can then be neglected. This echoes the conclusion obtained above in Section 4 that, at first order in ε, the effects of metric perturbations are confined to scales comparable to the variation scales of the perturbations.

The second order term $S_{G2}[n]$ cannot be computed exactly. Considering the conclusions of Section 4, one nevertheless expects increasing, possibly exponential functions of the time t to contribute to $S_{G2}[n]$, the characteristic time scale $T_{nn'}$ of these functions being related to the differences $k_n - k_{n'}$ in wave numbers of the metric perturbation by $T_{nn'} = 1/(\chi(k_n - k_{n'})^2)$. This expectation can be confirmed by computing exactly the contribution of $N_1 l_1$ to $S_{G2}[n]$. Naturally, given a certain wave number k, $T_{nn'}$ may be comparable to $T = 1/(\chi k^2)$, even if both $\mid k_n \mid$ and $\mid k_{n'} \mid$ are much larger than $\mid k \mid$. The behaviour of the Gibbs entropy thus confirms that, at second order, metric irregularities an influence diffusions at all scales, including scales much larger than the typical variation scales of the metric perturbation.

6. Conclusion

We have investigated how metric irregularities influence Brownian motion on a differential manifold. We have performed explicit perturbative calculations for nearly flat 2D manifolds and reached the conclusion that the metric irregularities have a cumulative effect on Brownian motion; more precisely, we have found that the relative difference of the spatial Fourier components of the densities generated by a Brownian motion on the flat surface and a Brownian motion on the irregular surface grows exponentially with time on all spatial scales, including scales much larger than those characteristic of the metric perturbation; entropy behavior has also been considered and characteristic times have been derived.

Let us conclude this article by mentioning some problems left open for further study. As stated in the introduction, many biological phenomena involve lateral diffusions on 2D interfaces. The results of this article suggest that the fluctuations of the interfaces profoundly affect these lateral diffusions; the discrepancies between real diffusions on irregular interfaces and idealized diffusions on highly regular surfaces are therefore probably observable and the biological consequences of these discrpancies should be carefully studied. On the theoretical side, one should envisage a non perturbative treatment of at least some of the problems studied in this article; this is probably best achieved through numerical simulations; a first step would be to confirm numerically, at least for 2D diffusions, the characteristic time estimates we have derived here. Finally, the case of relativistic diffusions in fluctuating space-times is certainly worth investigating, notably in a cosmological context.

References

[1] S. Abarbanel and A. Ditkowski, Asymptotically stable fourth-order accurate schemes for the diffusion equation on complex shapes, J. Comput. Phys., 133 (1996) 279.

[2] L. J. S. Allen, An Introduction to Stochastic Processes with Applications to Biology, Prentice Hall, 2003.

[3] J. Braga, J. M. P. Desterro and M. Carmo-Fonseca, Mol. Bio. Cell, 15 (2004) 4749-4760.

[4] A. Brünger, R. Peters and K. Schulten, Continuous fluorescence microphotolysis to observe lateral diffusion in membranes: theoretical methods and applications, J. Chem. Phys., 82 (1984) 2147.

[5] C. Chevalier and F. Debbasch, Fluctuation-Dissipation Theorems in an expanding Universe, J. Math. Phys., 48 (2007) 023304.

[6] M. Christensen, How to simulate anisotropic diffusion processes on curved surfaces, J. Comput. Phys., 201 (2004) 421-435.

[7] B. A. Dubrovin, S. P. Novikov and A. T. Fomenko, Modern geometry - Methods and applications, Springer-Verlag, New-York, 1984.

[8] L. Edelstein-Keshet, Mathematical Models in Biology, Classics in Applied Mathematics 46, SIAM, 2005.

[9] M. Emery, Stochastic calculus in manifolds, Springer-Verlag, 1989.

[10] N. S. Goel and N. Richter-Dyn, Stochastic Models in Biology, The Blackburn Press, 2004.

[11] N. Ikeda and S. Watanabe, Stochastic differential equations and diffusion processes, North-Holland Mathematical Library, 2nd edition, 1989.

[12] K. Itô, On stochastic differential equations on a differentiable manifold i., Nagoya Math. J., 1 (1950) 35-47.

[13] K. Itô, On stochastic differential equations on a differentiable manifold ii., MK, 28 (1953) 82-85.

[14] H. P. McKean, Stochastic integrals, Academic Press, New York and London, 1969.

[15] D. Mitra and Q. Wang, Stochastic traffic engineering for demand uncertainty and risk-aware network revenue management, IEEE/ACM Transactions on Networking, 13(2) (2005) 221-233.

[16] J. D. Murray, Mathematical Biology I: An Introduction, 3rd Edition, Interdisciplinary Applied Mathematics, Mathematical Biology, Springer, 2002.

[17] S. Nehls et al, Dynamics and retention of misfolded proteines in native er membranes, Nat. Cell. Bio., 2 (2000) 288-295.

[18] B. Øksendal, Stochastic Differential Equations, Universitext. Springer-Verlag, Berlin, 5th edition, 1998.

[19] I. F. Sbalzarini, A. Hayer, A. Helenius and P. Koumoutsakos, Simulations of (an)isotropic diffusion on curved biological interfaces. Biophysical J, 90(3) (2006) 878-885.

[20] M. Schreckenberg, A. Schadschneider, K. Nagel and N. Ito, Discrete stochastic models for traffic flow, Phys. Rev. E, 51(4) (1995) 2939-2949.

[21] S. E. Shreve, Stochastic Calculus for Finance I: The Binomial Asset Pricing Model, Springer Finance, Springer-Verlag, New-York, 2004.

[22] S. E. Shreve, Stochastic Calculus for Finance II: Continuous-Time Models, Springer Finance, Springer-Verlag, New-York, 2004.

[23] N. G. van Kampen, Stochastic Processes in Physics and Chemistry, North-Holland, Amsterdam, 1992.

Université Pierre et Marie Curie-Paris6, UMR 8112, ERGA-LERMA, 3 rue Galilée, 94200 Ivry, France.

E-mail: chevalier_claire@yahoo.fr and fabrice.debbasch@gmail.com

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 47–63

CONVERGENCE OF HIGH ORDER METHODS FOR MISCIBLE DISPLACEMENT

YEKATERINA EPSHTEYN AND BEATRICE RIVIÈRE

Abstract. We derive error estimates for a fully discrete scheme using primal discontinuous Galerkin discretization in space and backward Euler discretization in time. The estimates in the energy norm are optimal with respect to the mesh size and suboptimal with respect to the polynomial degree. The proposed scheme is of high order as polynomial approximations of pressure and concentration can take any degree. In addition, the method can handle different types of boundary conditions and is well-suited for unstructured meshes.

Key Words. flow, transport, porous media, miscible displacement, NIPG, SIPG, IIPG, h and p-version, fully discrete scheme.

1. Introduction

A high order numerical method for solving miscible displacement is introduced and analyzed in this paper. Miscible displacement occurs in important applications such as remediation of contaminated groundwater and production of oil from petroleum reservoirs. The physical model that describes the miscible displacement phenomena arises from the natural law of conservation of mass. This law is applied to each component of the fluid mixture. The mathematical model consists of a coupled system of partial differential equations: a pressure equation and a concentration equation for each component. Since the components of the fluid mixture may react with each other, the numerical method must accurately solve the laws of conservation. In particular, it is important to solve the continuity equation that describes the flow phenomena with high accuracy.

In this work, we propose a fully discrete scheme that is locally mass conservative. The approximations of pressure and concentration at each time step are discontinuous piecewise polynomials of different degrees. We show convergence of the numerical method with respect to both the mesh size and the polynomial degree. The flexibility inherent to discontinuous approximation spaces allows the use of complicated geometries and unstructured meshes. The primal discontinuous Galerkin method, analyzed in this paper, encompasses the nonsymmetric interior penalty Galerkin (NIPG) method, the symmetric interior penalty Galerkin (SIPG) and the incomplete interior penalty Galerkin (IIPG) method introduced for elliptic problems in [18, 26, 4]. Discontinuous Galerkin methods have been recently popular in modeling complex flow and transport problems in porous media (see for instance [22, 6, 5, 10, 14]).

Received by the editors September 24, 2007.

2000 *Mathematics Subject Classification.* 35Q35, 65N30, 65N15, 76S05.

The research of Y.Epshteyn is based upon work supported by the Center for Nonlinear Analysis (CNA) under the National Science Foundation Grant # DMS 0635983. This research was also partially supported by the National Science Foundation Grant # DMS 0506039. The research of B. Riviere was supported by the National Science Foundation Grant # DMS 0506039.

Several methods for solving the miscible displacement are proposed and analyzed in the literature. When classical continuous finite element approximations are used for both the pressure and the concentration equations, optimal convergence rates are proved in the dispersion-free case and nearly optimal convergence rates in the dispersion case, under somewhat idealized circumstances [8]. However, this procedure does not handle the transport-dominated problem arising from the concentration equation. Strong improvement in the accuracy of the approximation of the concentration is obtained by considering interior penalty Galerkin methods that can be based on continuous piecewise polynomial spaces [27] or on discontinuous piecewise polynomial spaces [11]. In this case, the pressure equation is solved with a continuous finite element method and penalty terms involving the jumps in the normal derivative are introduced in the concentration equation.

In the miscible displacement problem, only the velocity enters the equation for the concentration and therefore a natural procedure for solving the pressure equation is the locally mass conservative mixed finite element method. The concentration equation can be handled either by a continuous finite element method [12, 13] or by a modified method of characteristics, which combines the time derivative and the advection terms as a directional derivative [16, 24, 3]. In [23], a combination of a continuous finite element method and the method of characteristics for the concentration equation and a standard continuous finite element method for the pressure equation is used. As in the above cases, time stepping is done along the characteristics.

More recently, primal discontinuous Galerkin methods have been applied and analyzed for solving the miscible displacement problem using a semi-discrete approach. The system of equations is discretized in space only. A combined mixed method for the pressure equation with NIPG for concentration equation is studied in [20]. Both pressure and concentration are approximated by the NIPG method in [21, 17]. However, the convergence result in [21] is valid only if the boundary condition for pressure is a Neumann type. The numerical scheme presented in this paper, is fully discrete and valid for both Dirichlet and Neumann boundary conditions for the pressure and Dirichlet, Neuman and mixed boundary conditions for the concentration.

The outline of the paper is as follows. Section 2 contains the model problem and assumptions on the data. The coupled discontinuous Galerkin scheme is formulated in Section 3. Existence and convergence of the numerical solution are obtained in Section 4. Extensions of the scheme to several types of boundary conditions are presented in Section 5.

2. Model Problem and Notation

Consider the miscible displacement of one incompressible fluid by another in a porous medium $\Omega \subset \mathbb{R}^2$ and over the time interval $(0, T)$. Let p denote the pressure in the fluid mixture and let c denote the concentration (fraction volume) of the displaced fluid in the fluid mixture. The partial differential equations describing

this type of flow are:

$$-\nabla \cdot (\frac{K}{\mu(c)}\nabla p) = f_1, \quad \text{in} \quad \Omega \times (0,T), \tag{1}$$

$$u = -\frac{K}{\mu(c)}\nabla p, \quad \text{in} \quad \Omega \times (0,T), \tag{2}$$

$$\varphi \frac{\partial c}{\partial t} + \nabla \cdot (uc - D(u)\nabla c) = f_2, \quad \text{in} \quad \Omega \times (0,T), \tag{3}$$

subject to the following boundary conditions:

$$p = p_{\text{dir}} \quad \text{on} \quad \Gamma_{\text{D}} \times [0,T], \tag{4}$$

$$u \cdot n = u_{\text{dir}} \quad \text{on} \quad \Gamma_{\text{N}} \times [0,T], \tag{5}$$

$$c = c_{\text{dir}} \quad \text{on} \quad \partial\Omega \times [0,T], \tag{6}$$

where $\Gamma_{\text{D}} \cup \Gamma_{\text{N}}$ is a partition of the boundary $\partial\Omega$. Equation (1), referred to as the pressure equation, is coupled with equation (3) through the viscosity of the fluid mixture. Equation (3), referred to as the concentration equation, is coupled with equation (1) through the fluid velocity (2) and the dispersion-diffusion tensor $D(u)$:

$$D(u) = (\alpha_t \|u\|_2 + d_m)I + (\alpha_l - \alpha_t)\frac{uu^T}{\|u\|_2}.$$

The coefficient d_m is the molecular diffusivity, α_l and α_t are the longitudinal and transverse dispersivities, $\|u\|_2$ is the Euclidean norm of the velocity and I is the identity matrix. Let us also note, that the permeability K in the velocity equation (2) is obtained from a macroscopic averaging of the microscopic features of the medium. Hence, it can be discontinuous in space variable and can vary over several orders of magnitude. The coefficient ϕ in (3) is the porosity. Assumptions on the coefficients are made below.

Assumption H1. The function μ^{-1} is positive, bounded below and above by $\underline{\mu}$ and $\overline{\mu}$ respectively and it is also Lipschitz continuous.

$$\forall x_1, x_2 \in \mathbb{R}, \left|\frac{1}{\mu(x_1)} - \frac{1}{\mu(x_2)}\right| \leq \mu_L |x_1 - x_2|. \tag{7}$$

Assumption H2. The matrix K is symmetric positive definite and uniformly bounded above and below. There are positive constants $\overline{k}, \underline{k}$ such that:

$$\forall x \in \mathbb{R}^2, \quad \underline{k} x^T x \leq x^T K x \leq \overline{k} x^T x. \tag{8}$$

Assumption H3. The diffusion coefficient is strictly positive and the dispersivities are bounded.

$$\forall x \in \mathbb{R}^2, \quad 0 \leq \alpha_l(x) \leq \overline{\alpha_l}, \quad 0 \leq \alpha_t(x) \leq \overline{\alpha_t}, \text{ and } 0 < \underline{d} \leq d_m.$$

Under assumption H3 it was shown that $D(u)$ is uniformly positive definite in Ω and Lipschitz continuous [21]:

$$\forall u \in \mathbb{R}^2, \quad \forall x \in \mathbb{R}^2, \quad \underline{d} x^T x \leq x^T D(u) x, \tag{9}$$

$$\forall u, v \in \mathbb{R}^2, \quad \|D(u) - D(v)\|_2 \leq k_2 \|u - v\|_2, \tag{10}$$

Assumption H4. The matrix $D(u)$ is uniformly bounded above.

$$\forall u \in \mathbb{R}^2, \quad \forall x \in \mathbb{R}^2, x^T D(u) x \leq \overline{d} x^T x. \tag{11}$$

We propose a discontinuous finite element discretization of (1)-(6). For this, we introduce a non-degenerate quasi-uniform subdivision of Ω, made of either triangles or quadrilaterals. The quasi-uniformity assumption is only needed for the p-version,

i.e. for deriving error estimates in terms of the polynomial degree. As usual, the maximum diameter over all mesh elements is denoted by h. The set of interior edges is denoted by Γ_h. To each edge e in Γ_h, we associate a unit normal vector n_e. For a boundary edge, n_e is chosen so that it coincides with the outward normal. The space of discontinuous polynomials of degree $r \geq 1$ is denoted by $\mathcal{D}_r(\mathcal{E}_h)$:

$$\mathcal{D}_r(\mathcal{E}_h) = \{v \in L^2(\Omega) : \forall E \in \mathcal{E}_h : v|_E \in \mathbb{P}_r(E)\}.$$

For any function $v \in \mathcal{D}_r(\mathcal{E}_h)$, we denote the jump and average over a given edge e by $[v]$ and $\{v\}$ respectively. Assuming that n_e is outward to E_e^1, we can write:

$$\forall e = \partial E_e^1 \cap \partial E_e^2, \quad [v]|_e = v|_{E_e^1} - v|_{E_e^2}, \quad \{v\}|_e = 0.5v|_{E_e^1} + 0.5v|_{E_e^2},$$
$$\forall e = \partial E_e^1 \cap \partial\Omega, \quad [v]|_e = v|_{E_e^1}, \quad \{v\}|_e = v|_{E_e^1}.$$

Let N be a positive integer and let $\Delta t = T/N$ be the time step. Denote $t^i = i\Delta t$ for $0 \leq i \leq N$. Define the space

$$\mathcal{D}_{r,h}^N = \{\boldsymbol{v} = (v^i)_{0\leq i\leq N} : \quad \forall 0 \leq i \leq N \quad v^i \in \mathcal{D}_r(\mathcal{E}_h)\}.$$

We also denote by $\tilde{M}$ the constant that only depends on the maximum number of neighbors that one mesh element can have so that the following inequality holds. Let A be any quantity depending on E_e^1 or E_e^2:

$$\forall i = 1, 2, \quad (\sum_{e\in\Gamma_h} A(E_e^i))^{1/2} \leq \frac{\sqrt{\tilde{M}}}{2}(\sum_{E\in\mathcal{E}_h} A(E))^{1/2}. \tag{12}$$

$$(\sum_{e\in\Gamma_{\mathrm{D}}} A(E_e^1))^{1/2} \leq \sqrt{\tilde{M}}(\sum_{E\in\mathcal{E}_h} A(E))^{1/2}. \tag{13}$$

Let $H^k(\mathcal{O})$ be the usual Sobolev space on $\mathcal{O} \subset \mathbb{R}^d, d \geq 1$ with norm $\|\cdot\|_{k,\mathcal{O}}$. We also define the broken norm: $|||v|||_{k,\Omega} = \left(\sum_{E\in\mathcal{E}_h} \|v\|_{k,E}^2\right)^{1/2}$. We now recall well-known trace results and inverse inequality used in the error analysis [1, 25, 19].

Lemma 1. *There is a constant M_t independent of h such that if E is a triangle or quadrilateral, for any $e \subset \partial E$:*

$$\forall v \in H^s(E), s \geq 1, \|v\|_{0,e} \leq M_t h^{-1/2}(\|v\|_{0,E} + h\|\nabla v\|_{0,E}), \tag{14}$$

$$\forall v \in H^s(E), s \geq 2, \|\nabla v \cdot n\|_{0,e} \leq M_t h^{-1/2}(\|\nabla v\|_{0,E} + h\left\|\nabla^2 v\right\|_{0,E}). \tag{15}$$

Lemma 2. *Let E be a mesh element. Let $g : \mathbb{N} \to \mathbb{N}$ be a function defined by $g(k) = (k+1)(k+2)$ if E is a triangle, and by $g(k) = k^2$ if E is a quadrilateral. There is a constant M_t independent of h and k such that:*

$$\forall v \in \mathbb{P}_k(E), \forall e \subset \partial E, \|v\|_{0,e} \leq M_t\sqrt{\frac{g(k)}{h}}\|v\|_{0,E}. \tag{16}$$

Lemma 3 (Inverse Inequalities)**.** *Let E be a mesh element and $v \in \mathrm{P}_r(E)$. Then there exists a constant C independent of h and r such that*

$$\|v\|_{L^\infty(E)} \leq Ch^{-1}r^2\|v\|_{0,E}, \tag{17}$$

$$\|v\|_{1,E} \leq Ch^{-1}r^2\|v\|_{0,E}. \tag{18}$$

3. Scheme

At each discrete time t^i, we will approximate the pressure $p(t^i,\cdot)$ and concentration $c(t^i,\cdot)$ by discontinuous piecewise polynomials P^i and C^i of degree r_p and r_c respectively. For the p-version, we assume that the degrees are related in the following fashion. There exist positive constants δ_0, δ_1 such that

$$\delta_0 \leq \frac{r_c}{r_p} \leq \delta_1. \tag{19}$$

Before formulating the scheme, we introduce additional notation. Let ε be a parameter that takes the value -1, 0 or 1. By changing the value of ε, we will obtain the NIPG, SIPG or IIPG method. Let $\sigma_p > 0$ and $\sigma_c > 0$ be the penalty parameters. Our numerical method is the following: find $\boldsymbol{P} = (P^i)_{0\leq i\leq N} \in \mathcal{D}^N_{r_p,h}$ and $\boldsymbol{C} = (C^i)_{0\leq i\leq N} \in \mathcal{D}^N_{r_c,h}$ such that

Initial Concentration

$$\forall v \in \mathcal{D}_{r_c}(\mathcal{E}_h), \quad \int_\Omega C^0 v = \int_\Omega c^0 v. \tag{20}$$

Pressure Equation: $\forall 0 \leq i \leq N-1$,

$$\begin{aligned}
\forall z \in \mathcal{D}_{r_p}(\mathcal{E}_h), \quad & \sum_{E\in\mathcal{E}_h} \int_E \frac{1}{\mu(C^{i+1})} K\nabla P^{i+1}\cdot\nabla z + \sigma_p \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \frac{g(r_p)}{|e|}\int_e [P^{i+1}][z] \\
& - \sum_{e\in\Gamma_h}\int_e \{\frac{1}{\mu(C^{i+1})} K\nabla P^{i+1}\cdot n_e\}[z] - \sum_{e\in\Gamma_{\mathrm{D}}}\int_e \frac{1}{\mu(c_{\mathrm{dir}})} K\nabla P^{i+1}\cdot n_e z \\
& + \varepsilon\sum_{e\in\Gamma_h}\int_e \{\frac{1}{\mu(C^{i+1})} K\nabla z\cdot n_e\}[P^{i+1}] + \varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e \frac{1}{\mu(c_{\mathrm{dir}})} K\nabla z\cdot n_e P^{i+1} \\
= & \int_\Omega f_1 z + \sum_{e\in\Gamma_{\mathrm{N}}}\int_e u_{\mathrm{dir}} z + \sigma_p \sum_{e\in\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\int_e p_{\mathrm{dir}} z + \varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e \frac{1}{\mu(c_{\mathrm{dir}})} K\nabla z\cdot n_e p_{\mathrm{dir}}.
\end{aligned} \tag{21}$$

Concentration Equation: $\forall 0 \leq i \leq N-1$,

$$\begin{aligned}
\forall v \in \mathcal{D}_{r_c}(\mathcal{E}_h), \quad & \int_\Omega \frac{\varphi}{\Delta t}(C^{i+1}-C^i)v + \sum_{E\in\mathcal{E}_h}\int_E \frac{C^{i+1}}{\mu(C^{i+1})} K\nabla P^{i+1}\cdot\nabla v \\
& + \sum_{E\in\mathcal{E}_h}\int_E D(U^{i+1})\nabla C^{i+1}\cdot\nabla v - \sum_{e\in\Gamma_h}\int_e \{\frac{C^{i+1}}{\mu(C^{i+1})} K\nabla P^{i+1}\cdot n_e\}[v] \\
& - \sum_{e\in\Gamma_{\mathrm{D}}}\int_e \frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})} K\nabla P^{i+1}\cdot n_e v - \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\int_e \{D(U^{i+1})\nabla C^{i+1}\cdot n_e\}[v] \\
& + \varepsilon\sum_{e\in\Gamma_h}\int_e \{\frac{C^{i+1}}{\mu(C^{i+1})} K\nabla v\cdot n_e\}[P^{i+1}] + \varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e \frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})} K\nabla v\cdot n_e P^{i+1} \\
& + \varepsilon\sum_{e\in\Gamma_h}\int_e \{D(U^{i+1})\nabla v\cdot n_e\}[C^{i+1}] + \sigma_c \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\frac{g(r_c)}{|e|}\int_e [C^{i+1}][v]
\end{aligned} \tag{22}$$

$$= \varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e \frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})} K\nabla v\cdot n_e p_{\mathrm{dir}} + \sigma_c \sum_{e\in\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\frac{g(r_c)}{|e|}\int_e c_{\mathrm{dir}} v + \int_\Omega f_2 v + \sum_{e\in\Gamma_{\mathrm{N}}}\int_e c_{\mathrm{dir}} u_{\mathrm{dir}} v,$$

with the definition of the discrete velocity U^{i+1} given by

$$U^{i+1} = -\frac{K}{\mu(C^{i+1})}\nabla P^{i+1}. \tag{23}$$

We obtain a nonlinear system of equations that can be written in short as

$$\forall z \in \mathcal{D}_{r_p}(\mathcal{E}_h), \quad \forall v \in \mathcal{D}_{r_c}(\mathcal{E}_h), \quad 0 \le i \le N, \quad \mathcal{L}(P^i, C^i; z, v) = 0.$$

It is easy to check, using standard techniques for Interior Penalty discontinuous Galerkin methods, that the scheme (20)-(22) is consistent, i.e. if the solution of (1)-(6) is smooth enough and if we denote $c^i = c(t^i, \cdot)$ and $p^i = p(t^i, \cdot)$, then

$$\forall 0 \le i \le N, \quad \forall z \in \mathcal{D}_{r_p}(\mathcal{E}_h), \quad \forall v \in \mathcal{D}_{r_c}(\mathcal{E}_h), \quad \mathcal{L}(p^i, c^i; z, v) = 0, \tag{24}$$

4. Existence and Convergence of the Discrete Solution

In this section, we prove the existence and show convergence of the numerical solution $(\boldsymbol{P}, \boldsymbol{C})$ by the use of the Schauder's fixed point theorem (see for example theorem 6.44 in [9]). Let $\tilde{p}$ and $\tilde{c}$ be approximations of p and c. We assume that

$$\tilde{p} \in L^\infty(0, T, W^{1,\infty}(\Omega)), \quad \tilde{c} \in L^\infty(0, T, L^\infty(\Omega)), \quad \tilde{c}_{tt} \in L^\infty(0, T, L^2(\Omega)). \tag{25}$$

We will denote $\tilde{p}^i(\cdot) = \tilde{p}(t^i, \cdot)$ and $\tilde{c}^i(\cdot) = \tilde{c}(t^i, \cdot)$. We assume that there are constants $\kappa_p, \kappa_c \ge 2$ such that

$$\forall 0 \le i \le N, \quad \forall t > 0, \quad p^i(t) \in H^{\kappa_p}(\Omega), \quad c^i(t) \in H^{\kappa_c}(\Omega).$$

We assume that the following standard hp-type approximation results hold [2]

$$\forall 0 \le i \le N, \quad \|\tilde{p}^i - p^i\|_{H^s(\Omega)} \le M \frac{h^{\min(r_p+1, \kappa_p)-s}}{r_p^{\kappa_p - s}} \|p^i\|_{H^{\kappa_p}(\Omega)}, \tag{26}$$

$$\forall 0 \le i \le N, \quad \|\tilde{c}^i - c^i\|_{H^s(\Omega)} \le M \frac{h^{\min(r_c+1, \kappa_c)-s}}{r_c^{\kappa_c - s}} \|c^i\|_{H^{\kappa_c}(\Omega)}, \tag{27}$$

Here and throughout the paper, M is a generic constant independent of h, r_c, r_p and Δt, that takes different values at different places. In addition, in the case of the p-version, we assume that $\kappa_p, \kappa_c \ge 3$.
Next we prove existence and convergence of the solution using an idea similar to idea in [15]. Let us define the following subset of the broken Sobolev space:

$$\mathcal{W} = \Big\{ (\boldsymbol{\psi}, \boldsymbol{\phi}) \in \mathcal{D}^N_{r_p,h} \times \mathcal{D}^N_{r_c,h} : \phi^0 = \tilde{c}^0, \text{ and there exist positive constants}$$

$K_1, K_2, \ldots, K_6, \Delta t_0$ independent of h such that for $\Delta t \le \Delta t_0$ and $0 \le i \le N-1$:

$$\Big(\frac{1}{\Delta t} - K_1\Big) \left\|\phi^{i+1} - \widetilde{c}^{i+1}\right\|^2_{0,\Omega} - \frac{1}{\Delta t} \left\|\phi^i - \widetilde{c}^i\right\|^2_{0,\Omega}$$
$$+ |||\phi^{i+1} - \widetilde{c}^{i+1}|||_1^2 \le K_2 \frac{h^{2r_p}}{r_p^{2\kappa_p - 4}} + K_3 \frac{h^{2r_c}}{r_c^{2\kappa_c - 4}} + K_4 \Delta t^2,$$
$$|||\psi^{i+1} - \widetilde{p}^{i+1}|||_1^2 \le K_5 \frac{h^{2r_p}}{r_p^{2\kappa_p - 4}} + K_6 \frac{h^{2r_c}}{r_c^{2\kappa_c - 4}} \Big\}.$$

The set $\mathcal{W}$ is closed, convex subset of the broken Sobolev space and it is not empty since it contains the element $(\tilde{p}^i, \tilde{c}^i)_{0 \le i \le N}$.

Lemma 4. *For any $(\boldsymbol{\psi}, \boldsymbol{\phi})$ in $\mathcal{W}$, if Δt is small enough (namely $\Delta t = \mathcal{O}(h/r_c) < 1/K_1$), there exist positive constant M_1, M_2, M_3 for any $1 \le i \le N$*

$$\left\|\phi^i - \widetilde{c}^i\right\|_{0,\Omega} \le M_1 \Big(\frac{h^{r_p}}{r_p^{\kappa_p - 2}} + \frac{h^{r_c}}{r_c^{\kappa_c - 2}} + \Delta t \Big), \tag{28}$$

$$||\phi^i||_{\infty,\Omega} \le M_2, \quad |||\psi^i|||_1 \le M_3. \tag{29}$$

The constants M_1, M_2 are independent of h, r_p, r_c and Δt but depend on $K_1, \ldots, K_4$. The constant M_3 is independent of h, r_p, r_c and Δt but depends on K_5, K_6.

Proof. We will show that (29) is a consequence of the definition of $\mathcal{W}$. We first prove (28), which will yield (29). From the definition of the space $\mathcal{W}$, we have for $0 \le i \le N-1$:

$$\left\|\phi^{i+1}-\widetilde{c}^{i+1}\right\|_{0,\Omega}^2-\left\|\phi^{i}-\widetilde{c}^{i}\right\|_{0,\Omega}^2+\Delta t|||\phi^{i+1}-\widetilde{c}^{i+1}|||_1^2$$
$$\le \Delta t K_2\frac{h^{2r_p}}{r_p^{2\kappa_p-4}}+\Delta t K_3\frac{h^{2r_c}}{r_c^{2\kappa_c-4}}+K_4\Delta t^3+\Delta t K_1\left\|\phi^{i+1}-\widetilde{c}^{i+1}\right\|_{0,\Omega}^2.$$

Fix $n\ge 1$, sum from $i=0$ to $i=n-1$ and note that $\sum_{i=0}^{n-1}\Delta t\le T$ and $\phi^0=\tilde{c}^0$:

$$\|\phi^n-\widetilde{c}^n\|_{0,\Omega}^2+\Delta t\sum_{i=0}^{n-1}|||\phi^{i+1}-\widetilde{c}^{i+1}|||_1^2$$
$$\le K_2T\frac{h^{2r_p}}{r_p^{2\kappa_p-4}}+K_3T\frac{h^{2r_c}}{r_c^{2\kappa_c-4}}+K_4T\Delta t^2+\Delta t K_1\sum_{i=0}^{n-1}\left\|\phi^{i+1}-\widetilde{c}^{i+1}\right\|_{0,\Omega}^2.$$

From Gronwall's lemma, if $\Delta t<1/K_1$, there is a constant M independent of h and Δt such that

$$\|\phi^n-\widetilde{c}^n\|_{0,\Omega}^2+\Delta t\sum_{i=0}^{N-1}|||\phi^{i+1}-\widetilde{c}^{i+1}|||_1^2\le M(\frac{h^{2r_p}}{r_p^{2\kappa_p-4}}+\frac{h^{2r_c}}{r_c^{2\kappa_c-4}}+\Delta t^2).$$

This yields (28). Besides, from (19) and choosing $\Delta t=\mathcal{O}(\frac{h}{r_c^2})$, we conclude that

$$\|\phi^n-\widetilde{c}^n\|_{0,\Omega}\le M\frac{h}{r_c^2}.$$

Using an inverse inequality (17), we have

$$\|\phi^n-\widetilde{c}^n\|_{\infty,\Omega}\le Mr_c^2h^{-1}\|\phi^n-\widetilde{c}^n\|_{0,\Omega}\le M.$$

This implies that

$$\|\phi^n\|_{\infty,\Omega}\le M+\|\tilde{c}^n\|_{\infty,\Omega}\le M_2,$$

which with (25) yields gives the result (29). □

We now define an operator $\mathcal{F}$ that acts on elements in $\mathcal{W}$.

$$\forall(\boldsymbol{\psi},\boldsymbol{\phi})\in\mathcal{W},\quad \mathcal{F}(\boldsymbol{\psi},\boldsymbol{\phi})=(\boldsymbol{\psi}_L,\boldsymbol{\phi}_L),$$

where (ψ_L,ϕ_L) satisfies initial conditions:

$$(\psi_L^0,\phi_L^0)=(\psi^0,\phi^0), \tag{30}$$

and for $0\le i\le N-1$, $\psi_L^{i+1}\in\mathcal{D}_{r_p}(\mathcal{E}_h)$ and $\phi_L^{i+1}\in\mathcal{D}_{r_c}(\mathcal{E}_h)$ such that

$$\forall z\in\mathcal{D}_{r_p}(\mathcal{E}_h),\quad \sum_{E\in\mathcal{E}_h}\int_E\frac{1}{\mu(\phi^{i+1})}K\nabla\psi_L^{i+1}\cdot\nabla z+\sigma_p\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\int_e[\psi_L^{i+1}][z]$$
$$-\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(\phi^{i+1})}K\nabla\psi_L^{i+1}\cdot n_e\}[z]-\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla\psi_L^{i+1}\cdot n_ez$$
$$+\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(\phi^{i+1})}K\nabla z\cdot n_e\}[\psi_L^{i+1}]+\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla z\cdot n_e\psi_L^{i+1}$$
$$=\int_\Omega f_1z+\sum_{e\in\Gamma_{\mathrm{N}}}\int_e u_{\mathrm{dir}}z+\sigma_p\sum_{e\in\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\int_e p_{\mathrm{dir}}z+\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla z\cdot n_ep_{\mathrm{dir}}. \tag{31}$$

$$\forall v\in\mathcal{D}_{r_c}(\mathcal{E}_h),\quad \int_\Omega\frac{\varphi}{\Delta t}(\phi_L^{i+1}-\phi_L^i)v+\sum_{E\in\mathcal{E}_h}\int_E\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla\psi_L^{i+1}\cdot\nabla v$$

$$+\sum_{E\in\mathcal{E}_h}\int_E D(\zeta^{i+1})\nabla\phi_L^{i+1}\cdot\nabla v-\sum_{e\in\Gamma_h}\int_e\{\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla\psi_L^{i+1}\cdot n_e\}[v]$$

$$-\sum_{e\in\Gamma_{\rm D}}\int_e\frac{c_{\rm dir}}{\mu(c_{\rm dir})}K\nabla\psi_L^{i+1}\cdot n_e v-\sum_{e\in\Gamma_h\cup\Gamma_{\rm D}\cup\Gamma_{\rm N}}\int_e\{D(\zeta^{i+1})\nabla\phi_L^{i+1}\cdot n_e\}[v]$$

$$+\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla v\cdot n_e\}[\psi_L^{i+1}]+\varepsilon\sum_{e\in\Gamma_{\rm D}}\int_e\frac{c_{\rm dir}}{\mu(c_{\rm dir})}K\nabla v\cdot n_e\psi_L^{i+1}$$

$$+\varepsilon\sum_{e\in\Gamma_h}\int_e\{D(\zeta^{i+1})\nabla v\cdot n_e\}[\phi_L^{i+1}]+\sigma_c\sum_{e\in\Gamma_h\cup\Gamma_{\rm D}\cup\Gamma_{\rm N}}\frac{g(r_c)}{|e|}\int_e[\phi_L^{i+1}][v]$$

(32)

$$=\varepsilon\sum_{e\in\Gamma_{\rm D}}\int_e\frac{c_{\rm dir}}{\mu(c_{\rm dir})}K\nabla v\cdot n_e p_{\rm dir}+\sigma_c\sum_{e\in\Gamma_{\rm D}\cup\Gamma_{\rm N}}\frac{g(r_c)}{|e|}\int_e c_{\rm dir}v+\int_\Omega f_2 v+\sum_{e\in\Gamma_{\rm N}}\int_e c_{\rm dir}u_{\rm dir}v,$$

where

$$\zeta^{i+1}=-\frac{K}{\mu(\phi^{i+1})}\nabla\psi^{i+1}. \tag{33}$$

We show that $\mathcal{F}$ is well-defined by proving existence and uniqueness of $(\boldsymbol{\psi}_L,\boldsymbol{\phi}_L)$.

Lemma 5. *There exists a unique solution $(\boldsymbol{\psi}_L,\boldsymbol{\phi}_L)\in\mathcal{D}_{r_p,h}^N\times\mathcal{D}_{r_c}^N$ that satisfies (30)-(32).*

Proof. Since the problem (30)-(32) is linear and finite-dimensional, it suffices to show uniqueness of the solution. Let $(\boldsymbol{\psi}_{L1},\boldsymbol{\phi}_{L1})$ and $(\boldsymbol{\psi}_{L2},\boldsymbol{\phi}_{L2})$ be two solutions and let $(\bar{\boldsymbol{\psi}},\bar{\boldsymbol{\phi}})$ denote their differences. Then, the pair $(\bar{\boldsymbol{\psi}},\bar{\boldsymbol{\phi}})$ satisfies (30)-(32) with zero data $f_1=p_{\rm dir}=u_{\rm dir}=c_{\rm dir}=f_2=0$ and $\phi_L^i=0$. Clearly, we have $(\bar\psi^0,\bar\phi^0)=(0,0)$. Fix $i\in\{0,\dots,N-1\}$ and choose the test function $z=\bar\psi^{i+1}$ in (31).

$$|||\frac{1}{\mu(\phi^{i+1})^{1/2}}K^{1/2}\nabla\bar\psi^{i+1}|||_{0,\Omega}^2+\sigma_p\sum_{e\in\Gamma_h\cup\Gamma_{\rm D}}\frac{g(r_p)}{|e|}\|[\bar\psi^{i+1}]\|_{0,e}^2$$

$$-(1-\epsilon)\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(\phi^{i+1})}K\nabla\bar\psi^{i+1}\cdot n_e\}[\bar\psi^{i+1}]-(1-\epsilon)\sum_{e\in\Gamma_{\rm D}}\int_e\frac{1}{\mu(c_{\rm dir})}K\nabla\bar\psi^{i+1}\cdot n_e\bar\psi^{i+1}=0.$$

If $\epsilon=1$, we directly have that $\bar\psi^{i+1}=0$. Otherwise, using assumption H1 and trace and inverse inequalities, we can bound the last two terms of the left-hand side of the equation above by

$$\frac{1}{2}|||\frac{1}{\mu(\phi^{i+1})^{1/2}}K^{1/2}\nabla\bar\psi^{i+1}|||_{0,\Omega}^2+M\sum_{e\in\Gamma_h\cup\Gamma_{\rm D}}\frac{1}{|e|}\|[\bar\psi^{i+1}]\|_{0,e}^2,$$

which implies that $\bar\psi^{i+1}=0$ if the penalty value σ_p is large enough. Next, we choose the test function $v=\bar\phi^{i+1}$ in (32). A similar argument gives that $\bar\phi^{i+1}=0$ if the penalty σ_c is large enough. □

We now show that the range of $\mathcal{F}$ is included in the space $\mathcal{W}$. The same technique can be used to show that $\mathcal{F}$ is continuous.

Theorem 1.

$$\forall(\boldsymbol{\psi},\boldsymbol{\phi})\in\mathcal{W},\quad \mathcal{F}(\boldsymbol{\psi},\boldsymbol{\phi})\in\mathcal{W}.$$

Proof. Let $(\boldsymbol{\psi},\boldsymbol{\phi})\in\mathcal{W}$, $(\boldsymbol{\psi}_L,\boldsymbol{\phi}_L)=\mathcal{F}(\boldsymbol{\psi},\boldsymbol{\phi})$ and denote

$$\forall 0\le i\le N,\quad \tau^i=\psi_L^i-\tilde{p}^i,\theta^i=p^i-\tilde{p}^i,\xi^i=\phi_L^i-\tilde{c}^i,\chi^i=c^i-\tilde{c}^i.$$

From the consistency equations (24), we have

$$\begin{aligned}
&\sum_{E\in\mathcal{E}_h}\int_E\frac{1}{\mu(\phi^{i+1})}K\nabla\widetilde{p}^{i+1}\cdot\nabla z+\sigma_p\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\int_e[\widetilde{p}^{i+1}][z]\\
-\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(\phi^{i+1})}K\nabla\widetilde{p}^{i+1}\cdot n_e\}[z]-\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla\widetilde{p}^{i+1}\cdot n_e z-\sum_{e\in\Gamma_{\mathrm{N}}}\int_e u_{\mathrm{dir}}z\\
+\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(\phi^{i+1})}K\nabla z\cdot n_e\}[\widetilde{p}^{i+1}]+\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla z\cdot n_e\widetilde{p}^{i+1}\\
-\int_\Omega f_1z-\sigma_p\sum_{e\in\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\int_e p_{\mathrm{dir}}z-\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla z\cdot n_e p_{\mathrm{dir}}\\
=-\sum_{E\in\mathcal{E}_h}\int_E\frac{1}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot\nabla z-\sigma_p\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\int_e[\theta^{i+1}][z]\\
+\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot n_e\}[z]+\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla\theta^{i+1}\cdot n_e z\\
-\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(c^{i+1})}K\nabla z\cdot n_e\}[\theta^{i+1}]-\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla z\cdot n_e\theta^{i+1}\\
+\sum_{E\in\mathcal{E}_h}\int_E(\frac{1}{\mu(\phi^{i+1})}-\frac{1}{\mu(c^{i+1})})K\nabla\widetilde{p}^{i+1}\cdot\nabla z-\sum_{e\in\Gamma_h}\int_e\{(\frac{1}{\mu(\phi^{i+1})}-\frac{1}{\mu(c^{i+1})})K\nabla\widetilde{p}^{i+1}\cdot n_e\}[z]\\
+\varepsilon\sum_{e\in\Gamma_h}\int_e\{(\frac{1}{\mu(\phi^{i+1})}-\frac{1}{\mu(c^{i+1})})K\nabla z\cdot n_e\}[\widetilde{p}^{i+1}].
\end{aligned}\tag{34}$$

Subtracting equation (34) from (31) and choosing $z=\tau^{i+1}$, we obtain:

$$\begin{aligned}
&|||\frac{1}{\mu(\phi^{i+1})^{1/2}}K^{1/2}\nabla\tau^{i+1}|||_{0,\Omega}^2+\sigma_p\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\|[\tau^{i+1}]\|_{0,e}^2\\
=(1-\varepsilon)\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(\phi^{i+1})}K\nabla\tau^{i+1}\cdot n_e\}[\tau^{i+1}]+(1-\varepsilon)\sum_{e\in\Gamma_{\mathrm{D}}}\int_e(\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla\tau^{i+1}\cdot n_e)\tau^{i+1}\\
-\sum_{E\in\mathcal{E}_h}\int_E\frac{1}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot\nabla\tau^{i+1}-\sigma_p\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\frac{g(r_p)}{|e|}\int_e[\theta^{i+1}][\tau^{i+1}]\\
+\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot n_e\}[\tau^{i+1}]+\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla\theta^{i+1}\cdot n_e\tau^{i+1}\\
-\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{1}{\mu(c^{i+1})}K\nabla\tau^{i+1}\cdot n_e\}[\theta^{i+1}]-\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{1}{\mu(c_{\mathrm{dir}})}K\nabla\tau^{i+1}\cdot n_e\theta^{i+1}\\
+\sum_{E\in\mathcal{E}_h}\int_E(\frac{1}{\mu(\phi^{i+1})}-\frac{1}{\mu(c^{i+1})})K\nabla\widetilde{p}^{i+1}\cdot\nabla\tau^{i+1}\\
-\sum_{e\in\Gamma_h}\int_e\{(\frac{1}{\mu(\phi^{i+1})}-\frac{1}{\mu(c^{i+1})})K\nabla\widetilde{p}^{i+1}\cdot n_e\}[\tau^{i+1}]\\
+\varepsilon\sum_{e\in\Gamma_h}\int_e\{(\frac{1}{\mu(\phi^{i+1})}-\frac{1}{\mu(c^{i+1})})K\nabla z\cdot n_e\}[\widetilde{p}^{i+1}]=T_1+\cdots+T_{11}.
\end{aligned}\tag{35}$$

Next, we bound each term in the right-hand side of (35) using techniques standard for discontinuous Galerkin methods. In what follows, the quantities ε_i are positive real numbers to be defined later. Using Assumptions $H1$ and $H2$ and Cauchy-Schwarz inequality, we have

$$|T_1| \leq (1-\varepsilon)\overline{\mu} \sum_{e\in\Gamma_h} \left\|\{K^{\frac{1}{2}}\nabla\tau^{i+1}\}\right\|_{0,e} \left\|[\tau^{i+1}]\right\|_{0,e}$$

We now fix an interior edge e and denote E_e^1 and E_e^2 two elements sharing the edge e. Using (12) and the trace inequality (16), we have:

$$\sum_{e\in\Gamma_h} \left\|\{K^{\frac{1}{2}}\nabla\tau^{i+1}\}\right\|_{0,e} \left\|[\tau^{i+1}]\right\|_{0,e} \leq \sum_{e\in\Gamma_h} \frac{1}{2}(\left\|K^{\frac{1}{2}}\nabla\tau^{i+1}|_{E_e^1}\right\|_{0,e} + \left\|K^{\frac{1}{2}}\nabla\tau^{i+1}|_{E_e^2}\right\|_{0,e}) \left\|[\tau^{i+1}]\right\|_{0,e}$$

$$\leq \frac{1}{2}M_t\sqrt{\frac{g(r_p)}{h}} \sum_{e\in\Gamma_h} (\left\|K^{\frac{1}{2}}\nabla\tau^{i+1}\right\|_{0,E_e^1} + \left\|K^{\frac{1}{2}}\nabla\tau^{i+1}\right\|_{0,E_e^2}) \left\|[\tau^{i+1}]\right\|_{0,e}$$

$$\leq (\sum_{e\in\Gamma_h} \frac{M_t^2 g(r_p)}{4h} \left\|[\tau^{i+1}]\right\|_{0,e}^2)^{\frac{1}{2}} \Big((\sum_{e\in\Gamma_h} \left\|K^{\frac{1}{2}}\nabla\tau^{i+1}\right\|_{0,E_e^1}^2)^{\frac{1}{2}} + (\sum_{e\in\Gamma_h} \left\|K^{\frac{1}{2}}\nabla\tau^{i+1}\right\|_{0,E_e^2}^2)^{\frac{1}{2}} \Big)$$

$$\leq (\sum_{e\in\Gamma_h} \frac{\widetilde{M}M_t^2 g(r_p)}{4h} \left\|[\tau^{i+1}]\right\|_{0,e}^2)^{\frac{1}{2}} (\sum_{E\in\mathcal{E}_h} \left\|K^{\frac{1}{2}}\nabla\tau^{i+1}\right\|_{0,E}^2)^{\frac{1}{2}}.$$

Therefore, we have the following bound on T_1:

$$|T_1| \leq \frac{\underline{\mu}}{24}|||K^{\frac{1}{2}}\nabla\tau^{i+1}|||_{0,\Omega}^2 + (1-\varepsilon)^2\frac{3(\overline{\mu})^2\overline{k}\widetilde{M}M_t^2}{2\underline{\mu}} \sum_{e\in\Gamma_h} \frac{g(r_p)}{h} \left\|[\tau^{i+1}]\right\|_{0,e}^2.$$

Similarly, using (13) and (16), we have for T_2:

$$|T_2| \leq \frac{\underline{\mu}}{24}|||K^{\frac{1}{2}}\nabla\tau^{i+1}|||_{0,\Omega}^2 + (1-\varepsilon)^2\frac{6(\overline{\mu})^2\overline{k}\widetilde{M}M_t^2}{\underline{\mu}} \sum_{e\in\Gamma_{\mathrm{D}}} \frac{g(r_p)}{|e|} \left\|[\tau^{i+1}]\right\|_{0,e}^2.$$

The term T_3 is bounded using assumption H1 and (8), Cauchy-Schwarz and Young's inequalities:

$$|T_3| \leq \frac{\underline{\mu}}{12}|||K^{\frac{1}{2}}\nabla\tau^{i+1}|||_{0,\Omega}^2 + M|||\nabla\theta^{i+1}|||_{0,\Omega}^2.$$

Using the trace inequality (14), we have for the term T_4:

$$|T_4| \leq \frac{\sigma_p}{8} \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \frac{g(r_p)}{|e|} \left\|[\tau^{i+1}]\right\|_{0,e}^2 + Mg(r_p) \sum_{E\in\mathcal{E}_h} (h^{-2}\left\|\theta^{i+1}\right\|_{0,E}^2 + \left\|\nabla\theta^{i+1}\right\|_{0,E}^2).$$

The terms T_5 and T_6 are bounded in a similar way as the terms T_1 and T_2, except that the trace inequality (15) is used instead of (16).

$$|T_5|+|T_6| \leq \frac{\sigma_p}{8} \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \frac{g(r_p)}{|e|} \left\|[\tau^{i+1}]\right\|_{0,e}^2 + \frac{M}{g(r_p)} \sum_{E\in\mathcal{E}_h} (\left\|\nabla\theta^{i+1}\right\|_{0,E}^2 + h^2\left\|\nabla^2\theta^{i+1}\right\|_{0,E}^2).$$

The terms T_7 and T_8 are handled in the same way as terms T_1 and T_2, with the exception that the trace inequality (14) is used to handle the approximation error term.

$$|T_7| + |T_8| \leq \frac{\underline{\mu}}{6}|||K^{1/2}\nabla\tau^{i+1}|||_{0,\Omega}^2 + Mg(r_p) \sum_{E\in\mathcal{E}_h} (h^{-2}\left\|\theta^{i+1}\right\|_{0,E}^2 + \left\|\nabla\theta^{i+1}\right\|_{0,E}^2).$$

Using (7), (8), Cauchy-Schwarz inequality and assumption on $\widetilde{p}^{i+1}$ (25), we have:

$$|T_9| \leq \frac{\underline{\mu}}{12}|||K^{\frac{1}{2}}\nabla\tau^{i+1}|||_{0,\Omega}^2 + M\|\nabla\widetilde{p}^{i+1}\|_\infty^2\|\phi^{i+1}-\widetilde{c}^{i+1}\|_{0,\Omega}^2 + M\|\nabla\widetilde{p}^{i+1}\|_\infty^2 \left\|\chi^{i+1}\right\|_{0,\Omega}^2.$$

The term T_{10} is a summation term over interior edges. We assume that the edge e is shared by the elements E_e^1 and E_e^2. Thus, we have using (7), (8), (25) and Cauchy-Schwarz inequality:

$$|T_{10}| \le \left\|\nabla \widetilde{p}^{i+1}\right\|_\infty \overline{k}\frac{\mu_L}{2}\sum_{e\in\Gamma_h}\Big((\left\|(\phi^{i+1}-\widetilde{c}^{i+1})|_{E_e^1}\right\|_{0,e}+\left\|(\phi^{i+1}-\widetilde{c}^{i+1})|_{E_e^2}\right\|_{0,e})\left\|[\tau^{i+1}]\right\|_{0,e}$$

$$+(\left\|\chi^{i+1}|_{E_e^1}\right\|_{0,e}+\left\|\chi^{i+1}|_{E_e^2}\right\|_{0,e})\left\|[\tau^{i+1}]\right\|_{0,e}\Big).$$

Using the trace inequality (14), (16), we have:

$$|T_{10}| \le \frac{\sigma_p}{8}\sum_{e\in\Gamma_h}\frac{g(r_c)}{|e|}\left\|[\tau^{i+1}]\right\|_{0,e}^2 + M\|\nabla\widetilde{p}^{i+1}\|_\infty^2\|\phi^{i+1}-\widetilde{c}^{i+1}\|_{0,\Omega}^2$$

$$+\frac{M\left\|\nabla\widetilde{p}^{i+1}\right\|_\infty^2}{g(r_c)}\sum_{E\in\mathcal{E}_h}(\left\|\chi^{i+1}\right\|_{0,E}^2+h^2\left\|\nabla\chi^{i+1}\right\|_{0,E}^2).$$

The term T_{11} vanishes if the approximation $\widetilde{p}$ is continuous. Otherwise, we can bound exactly like the term T_5.

$$|T_{11}| \le \frac{\underline{\mu}}{12}|||K^{1/2}\nabla\tau^{i+1}|||_{0,\Omega}^2 + Mg(r_p)\sum_{E\in\mathcal{E}_h}(h^{-2}\left\|\theta^{i+1}\right\|_{0,E}^2+\left\|\nabla\theta^{i+1}\right\|_{0,E}^2).$$

Combining all the bounds above, using the fact that $1\le r^2\le g(r)\le 6r^2$, we have the pressure:

$$\frac{\underline{\mu}}{2}|||K^{\frac{1}{2}}\nabla\tau^{i+1}|||_{0,\Omega}^2+(\frac{\sigma_p}{2}-(1-\varepsilon)^2\frac{3(\overline{\mu})^2\overline{k}\widetilde{M}M_t^2}{2\underline{\mu}})\sum_{e\in\Gamma_h}\frac{g(r_p)}{|e|}\left\|[\tau^{i+1}]\right\|_{0,e}^2$$

$$+(\frac{7}{8}\sigma_p-(1-\varepsilon)^2\frac{6(\overline{\mu})^2\overline{k}\widetilde{M}M_t^2}{\underline{\mu}})\sum_{e\in\Gamma_D}\frac{g(r_p)}{|e|}\left\|[\tau^{i+1}]\right\|_{0,e}^2 \le M\frac{g(r_p)}{h^2}\left\|\theta^{i+1}\right\|_{0,\Omega}^2$$

$$+M(1+g(r_p))|||\nabla\theta^{i+1}|||_{0,\Omega}^2+M\left\|\chi^{i+1}\right\|_{0,\Omega}^2+\frac{Mh^2}{g(r_c)}|||\nabla\chi^{i+1}|||_{0,\Omega}^2+M\|\phi^{i+1}-\widetilde{c}^{i+1}\|_{0,\Omega}^2.$$

Define the limiting value of the penalty parameter:

$$\sigma_p^* = (1-\varepsilon)^2\frac{48(\overline{\mu})^2\overline{k}\widetilde{M}M_t^2}{7\underline{\mu}}.$$

Assuming that $\sigma_p > \sigma_p^*$, using the approximation results and the fact that $\boldsymbol{\phi}$ belongs to $\mathcal{W}$, we obtain:

$$|||\nabla\tau^{i+1}|||_{0,\Omega}^2+\sum_{e\in\Gamma_h\cup\Gamma_D}\frac{g(r_p)}{|e|}\left\|[\tau^{i+1}]\right\|_{0,e}^2$$

$$\le M\max(\frac{1}{\underline{\mu}\underline{k}},\frac{1}{\sigma_p-\sigma_p^*})\Big(\frac{h^{2\min(r_p+1,\kappa_p)-2}}{r_p^{2\kappa_p-4}}\|p^{i+1}\|_{H^{\kappa_p}(\Omega)}^2$$

$$+\frac{h^{2\min(r_c+1,\kappa_c)-2}}{r_c^{2\kappa_c-2}}\|c^{i+1}\|_{H^{\kappa_c}(\Omega)}^2+M_1(\frac{h^{2r_p}}{r_p^{2\kappa_p-4}}+\frac{h^{2r_c}}{r_c^{2\kappa_c-4}}+\Delta t^2)\Big). \tag{36}$$

Next, we consider the concentration equation in the system (24). The same way as for the pressure equation, the concentration equation can be rewritten as:

$$\int_\Omega\frac{\phi}{\Delta t}(\widetilde{c}^{i+1}-\widetilde{c}^i)v+\sum_{E\in\mathcal{E}_h}\int_E\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla\widetilde{p}^{i+1}\cdot\nabla v+\sum_{E\in\mathcal{E}_h}\int_E D(\zeta^{i+1})\nabla\widetilde{c}^{i+1}\cdot\nabla v$$

$$
\begin{aligned}
&-\sum_{e\in\Gamma_h}\int_e\{\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla\widetilde{p}^{i+1}\cdot n_e\}[v]-\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})}K\nabla\widetilde{p}^{i+1}\cdot n_e v-\sum_{e\in\Gamma_{\mathrm{N}}}\int_e c_{\mathrm{dir}}u_{\mathrm{dir}}v\\
&\quad-\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\int_e\{D(\zeta^{i+1})\nabla\widetilde{c}^{i+1}\cdot n_e\}[v]+\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla v\cdot n_e\}[\widetilde{p}^{i+1}]\\
&\qquad+\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})}K\nabla v\cdot n_e\widetilde{p}^{i+1}+\varepsilon\sum_{e\in\Gamma_h}\int_e\{D(\zeta^{i+1})\nabla v\cdot n_e\}[\widetilde{c}^{i+1}]\\
&\qquad+\sigma_c\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\frac{g(r_c)}{|e|}\int_e[\widetilde{c}^{i+1}][v]-\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})}K\nabla v\cdot n_e p_{\mathrm{dir}}\\
&\qquad-\sigma_c\sum_{e\in\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\frac{g(r_c)}{|e|}\int_e c_{\mathrm{dir}}v-\int_\Omega f_2 v=\int_\Omega\phi\Delta t\rho^{i+1}v-\int_\Omega\phi\chi_t^{i+1}v\\
&\quad\qquad-\sum_{E\in\mathcal{E}_h}\int_E\frac{c^{i+1}}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot\nabla v-\sum_{E\in\mathcal{E}_h}\int_E D(u^{i+1})\nabla\chi^{i+1}\cdot\nabla v\\
&\quad\qquad+\sum_{e\in\Gamma_h}\int_e\{\frac{c^{i+1}}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot n_e\}[v]+\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})}K\nabla\theta^{i+1}\cdot n_e v\\
&\quad+\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\int_e\{D(u^{i+1})\nabla\chi^{i+1}\cdot n_e\}[v]-\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{c^{i+1}}{\mu(c^{i+1})}K\nabla v\cdot n_e\}[\theta^{i+1}]\\
&\qquad-\varepsilon\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})}K\nabla v\cdot n_e\theta^{i+1}-\varepsilon\sum_{e\in\Gamma_h}\int_e\{D(u^{i+1})\nabla v\cdot n_e\}[\chi^{i+1}]\\
&\quad-\sigma_c\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\frac{g(r_c)}{|e|}\int_e[\chi^{i+1}][v]+\sum_{E\in\mathcal{E}_h}\int_E(D(\zeta^{i+1})-D(u^{i+1}))\nabla\widetilde{c}^{i+1}\cdot\nabla v\\
&\qquad\qquad+\sum_{E\in\mathcal{E}_h}\int_E\Big(\frac{\phi^{i+1}}{\mu(\phi^{i+1})}-\frac{c^{i+1}}{\mu(c^{i+1})}\Big)K\nabla\widetilde{p}^{i+1}\cdot\nabla v\\
&\qquad\qquad-\sum_{e\in\Gamma_h}\int_e\{\Big(\frac{\phi^{i+1}}{\mu(\phi^{i+1})}-\frac{c^{i+1}}{\mu(c^{i+1})}\Big)K\nabla\widetilde{p}^{i+1}\cdot n_e\}[v]\\
&\qquad\qquad-\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\int_e\{\Big(D(\zeta^{i+1})-D(u^{i+1})\Big)\nabla\widetilde{c}^{i+1}\cdot n_e\}[v]\\
&\qquad\qquad+\varepsilon\sum_{e\in\Gamma_h}\int_e\{\Big(\frac{\phi^{i+1}}{\mu(\phi^{i+1})}-\frac{c^{i+1}}{\mu(c^{i+1})}\Big)K\nabla v\cdot n_e\}[\widetilde{p}^{i+1}]\\
&\qquad\qquad+\varepsilon\sum_{e\in\Gamma_h}\int_e\{\Big(D(\zeta^{i+1})-D(u^{i+1})\Big)\nabla v\cdot n_e\}[\widetilde{c}^{i+1}].
\end{aligned}
\tag{37}
$$

Subtracting equation (37) from (32), using (33) and choosing $z=\xi^{i+1}$, we obtain:

$$
\begin{aligned}
&\int_\Omega\frac{\phi}{\Delta t}(\xi^{i+1}-\xi^i)\xi^{i+1}+|||D(\zeta^{i+1})^{1/2}\nabla\xi^{i+1}|||_{0,\Omega}^2+\sigma_c\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\frac{g(r_c)}{|e|}\|[\xi^{i+1}]\|_{0,e}^2\\
&\quad=\sum_{E\in\mathcal{E}_h}\int_E\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla\tau^{i+1}\cdot\nabla\xi^{i+1}-\sum_{e\in\Gamma_h}\int_e\{\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla\tau^{i+1}\cdot n_e\}[\xi^{i+1}]\\
&\quad-\sum_{e\in\Gamma_{\mathrm{D}}}\int_e\frac{c_{\mathrm{dir}}}{\mu(c_{\mathrm{dir}})}K\nabla\tau^{i+1}\cdot n_e\xi^{i+1}+(1-\varepsilon)\sum_{e\in\Gamma_h}\int_e\{D(\zeta^{i+1})\nabla\xi^{i+1}\cdot n_e\}[\xi^{i+1}]
\end{aligned}
$$

$$+\sum_{e\in\Gamma_{\rm D}\cup\Gamma_{\rm N}}\int_e D(\zeta^{i+1})\nabla\xi^{i+1}\cdot n_e\xi^{i+1}-\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{\phi^{i+1}}{\mu(\phi^{i+1})}K\nabla\xi^{i+1}\cdot n_e\}[\tau^{i+1}]$$

$$-\varepsilon\sum_{e\in\Gamma_{\rm D}}\int_e\frac{c_{\rm dir}}{\mu(c_{\rm dir})}K\nabla\xi^{i+1}\cdot n_e\tau^{i+1}-\int_\Omega\phi\Delta t\rho^{i+1}\xi^{i+1}+\int_\Omega\phi\chi_t^{i+1}\xi^{i+1}$$

$$+\sum_{E\in\mathcal{E}_h}\int_E\frac{c^{i+1}}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot\nabla\xi^{i+1}+\sum_{E\in\mathcal{E}_h}\int_E D(u^{i+1})\nabla\chi^{i+1}\cdot\nabla\xi^{i+1}$$

$$-\sum_{e\in\Gamma_h}\int_e\{\frac{c^{i+1}}{\mu(c^{i+1})}K\nabla\theta^{i+1}\cdot n_e\}[\xi^{i+1}]-\sum_{e\in\Gamma_{\rm D}}\int_e\frac{c_{\rm dir}}{\mu(c_{\rm dir})}K\nabla\theta^{i+1}\cdot n_e\xi^{i+1}$$

$$-\sum_{e\in\Gamma_h\cup\Gamma_{\rm D}\cup\Gamma_{\rm N}}\int_e\{D(u^{i+1})\nabla\chi^{i+1}\cdot n_e\}[\xi^{i+1}]+\varepsilon\sum_{e\in\Gamma_h}\int_e\{\frac{c^{i+1}}{\mu(c^{i+1})}K\nabla\xi^{i+1}\cdot n_e\}[\theta^{i+1}]$$

$$+\varepsilon\sum_{e\in\Gamma_{\rm D}}\int_e\frac{c_{\rm dir}}{\mu(c_{\rm dir})}K\nabla\xi^{i+1}\cdot n_e\theta^{i+1}+\varepsilon\sum_{e\in\Gamma_h}\int_e\{D(u^{i+1})\nabla\xi^{i+1}\cdot n_e\}[\chi^{i+1}]$$

$$+\sigma_c\sum_{e\in\Gamma_h\cup\Gamma_{\rm D}\cup\Gamma_{\rm N}}\frac{g(r_c)}{|e|}\int_e[\chi^{i+1}][\xi^{i+1}]-\sum_{E\in\mathcal{E}_h}\int_E(D(\zeta^{i+1})-D(u^{i+1}))\nabla\widetilde{c}^{i+1}\cdot\nabla\xi^{i+1}$$

$$-\sum_{E\in\mathcal{E}_\langle}\int_E\Big(\frac{\phi^{i+1}}{\mu(\phi^{i+1})}-\frac{c^{i+1}}{\mu(c^{i+1})}\Big)K\nabla\widetilde{p}^{i+1}\cdot\nabla\xi^{i+1}$$

$$+\sum_{e\in\Gamma_h}\int_e\{\Big(\frac{\phi^{i+1}}{\mu(\phi^{i+1})}-\frac{c^{i+1}}{\mu(c^{i+1})}\Big)K\nabla\widetilde{p}^{i+1}\cdot n_e\}[\xi^{i+1}]$$

$$+\sum_{e\in\Gamma_h\cup\Gamma_{\rm D}\cup\Gamma_{\rm N}}\int_e\{\Big(D(\zeta^{i+1})-D(u^{i+1})\Big)\nabla\widetilde{c}^{i+1}\cdot n_e\}[\xi^{i+1}]$$

$$-\varepsilon\sum_{e\in\Gamma_h}\int_e\{\Big(\frac{\phi^{i+1}}{\mu(\phi^{i+1})}-\frac{c^{i+1}}{\mu(c^{i+1})}\Big)K\nabla\xi^{i+1}\cdot n_e\}[\widetilde{p}^{i+1}]$$

$$-\varepsilon\sum_{e\in\Gamma_h}\int_e\{\Big(D(\zeta^{i+1})-D(u^{i+1})\Big)\nabla\xi^{i+1}\cdot n_e\}[\widetilde{c}^{i+1}]=S_1+\ldots S_{24}. \tag{38}$$

The term S_8 contains the numerical error in the time discretization:

$$\rho^{i+1}=\frac{1}{\Delta t}(\frac{\tilde{c}^{i+1}-\tilde{c}^i}{\Delta t}-\frac{\partial\tilde{c}^{i+1}}{\partial t}).$$

We now bound each term in the right-hand side of (38). The terms $S_1,\ldots,S_{18}$ are bounded like the terms T_i's. We skip the details (see [7]). Consider the term S_{19} using the assumptions H1, H3 and that $(\boldsymbol{\psi},\boldsymbol{\phi})\in\mathcal{W}$ we have:

$$S_{19}\leq\frac{d}{28}|||\nabla\xi^{i+1}|||_0^2+M\|\nabla\tilde{c}^{i+1}\|_\infty^2(|||\nabla(\psi^{i+1}-\tilde{p}^{i+1})|||_0^2+|||\nabla\theta^{i+1}|||_0^2)$$
$$+M\|\nabla\tilde{c}^{i+1}\|_\infty^2\|\nabla p^{i+1}\|_\infty^2(\|\phi^{i+1}-\tilde{c}^{i+1}\|_0^2+\|\chi^{i+1}\|_0^2).$$

Before bounding the term S_{20} we remark that

$$|\frac{\phi^{i+1}}{\mu(\phi^{i+1})}-\frac{c^{i+1}}{\mu(c^{i+1})}|\leq\bar{\mu}^2(\frac{1}{\bar{\mu}}+\mu_L\|c^{i+1}\|_\infty)|\phi^{i+1}-c^{i+1}|.$$

Therefore we have

$$S_{20}\leq\frac{d}{28}|||\nabla\xi^{i+1}|||_0^2+M\|\nabla\tilde{p}^{i+1}\|_\infty^2(1+\|c^{i+1}\|_\infty)^2(\|\phi^{i+1}-\tilde{c}^{i+1}\|_0^2+\|\chi^{i+1}\|_0^2).$$

The term S_{21} is bounded similarly to the term T_{10}.Consider the term S_{22}, using the assumptions H1 and H3 we have:

$$S_{22} \leq \frac{\sigma_c}{18} \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}} \frac{g(r_c)}{|e|}\|[\xi^{i+1}]\|_{0,e}^2 + \frac{Mg(r_p)\|\nabla\tilde{c}^{i+1}\|_\infty^2}{g(r_c)}|||\nabla(\psi^{i+1}-\tilde{p}^{i+1})|||_1^2$$
$$+\frac{M\|\nabla\tilde{c}^{i+1}\|_\infty^2}{g(r_c)} \sum_{E\in\mathcal{E}_h} (\|\nabla\theta^{i+1}\|_E^2 + h^2\|\nabla^2\theta^{i+1}\|_E^2) + M\|\nabla\tilde{c}^{i+1}\|_\infty^2\|\nabla p^{i+1}\|_\infty^2\|\phi^{i+1}-\tilde{c}^{i+1}\|_0^2$$
$$+\frac{M\|\nabla\tilde{c}^{i+1}\|_\infty^2\|\nabla p^{i+1}\|_\infty^2}{g(r_c)} \sum_{E\in\mathcal{E}_h} (\|\chi^{i+1}\|_E^2 + h^2\|\nabla\chi^{i+1}\|_E^2).$$

The terms S_{23} and S_{24} are bounded like the term T_{11}. Combining the bounds above, using (36) and the fact that $1 \leq r^1 \leq g(r) \leq 6r^2$, we obtain the following estimate:

$$\frac{1}{2\Delta t}\Big(\Big\|\phi^{1/2}\xi^{i+1}\Big\|_{0,\Omega}^2 - \Big\|\phi^{1/2}\xi^{i}\Big\|_{0,\Omega}^2\Big) + \frac{\underline{d}}{2}|||\nabla\xi^{i+1}|||_{0,\Omega}^2$$
$$+\Big(\frac{\sigma_c}{3} - (1-\varepsilon)^2\frac{7(\overline{d})^2\widetilde{M}M_t^2}{4\underline{d}}\Big) \sum_{e\in\Gamma_h} \frac{g(r_c)}{|e|}\left\|[\xi^{i+1}]\right\|_{0,e}^2$$
$$+\Big(\frac{\sigma_c}{3} - \frac{7(\overline{d})^2\widetilde{M}M_t^2}{\underline{d}}\Big) \sum_{e\in\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}} \frac{g(r_c)}{h}\left\|[\xi^{i+1}]\right\|_{0,e}^2$$
$$\leq M|||\nabla\tau^{i+1}|||_0^2 + M\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}} \frac{g(r_p)}{|e|}\left\|[\tau^{i+1}]\right\|_{0,e}^2 + \frac{1}{2}\left\|\xi^{i+1}\right\|_{0,\Omega}^2 + M\Delta t^2\left\|\rho^{i+1}\right\|_{0,\Omega}^2$$
$$+M\left\|\chi_t^{i+1}\right\|_{0,\Omega}^2 + M(\frac{g(r_c)}{h^2}+1)\|\chi^{i+1}\|_{0,\Omega}^2 + M\frac{g(r_c)}{h^2}\|\theta^{i+1}\|_{0,\Omega}^2$$
$$+M(1+g(r_c))|||\nabla\chi^{i+1}|||_{0,\Omega}^2 + M(1+g(r_c))|||\nabla\theta^{i+1}|||_{0,\Omega}^2 + M\|\phi^{i+1}-\tilde{c}^{i+1}\|_{0,\Omega}^2$$
$$+M\frac{h^2}{g(r_c)}|||\nabla^2\chi^{i+1}|||_{0,\Omega}^2 + M\frac{h^2}{g(r_c)}|||\nabla^2\theta^{i+1}|||_{0,\Omega}^2 + M|||\nabla(\tilde{\psi}^{i+1}-\tilde{p}^{i+1})|||_{0,\Omega}^2.$$

The error $\|\rho^{i+1}\|_{0,\Omega}$ is bounded using a Taylor expansion with integral remainder:

$$\tilde{c}^i = \tilde{c}^{i+1} - \Delta t\frac{\partial\tilde{c}^{i+1}}{\partial t} + \frac{1}{2}\int_{t^i}^{t^{i+1}}(t-t^i)\frac{\partial^2\tilde{c}^{i+1}}{\partial t^2}dt,$$

which yields

$$\|\rho^{i+1}\|_{0,\Omega} \leq M\|\tilde{c}_{tt}\|_{L^\infty(t^i,t^{i+1},L^2(\Omega))}.$$

Define

$$\sigma_c^* = \max\Big((1-\varepsilon)^2\frac{21(\overline{d})^2\widetilde{M}M_t^2}{4\underline{d}}, \frac{21(\overline{d})^2\widetilde{M}M_t^2}{\underline{d}}\Big).$$

Under the condition $\sigma_c > \sigma_c^*$ and using the approximation result, we obtain the following estimate:

$$\frac{\left\|\varphi^{1/2}\xi^{i+1}\right\|_{0,\Omega}^2}{\Delta t} - \frac{\left\|\varphi^{1/2}\xi^{i}\right\|_{0,\Omega}^2}{\Delta t} + |||\nabla\xi^{i+1}|||_\Omega^2 + \sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}} \frac{g(r_c)}{|e|}\left\|[\xi^{i+1}]\right\|_{0,e}$$

$$\leq \max(1,\frac{1}{\underline{d}},\frac{3}{2(\sigma_c-\sigma_*)})\Big(\left\|\xi^{i+1}\right\|_{0,\Omega}^2 + K_2\frac{h^{2r_p}}{r_p^{2\kappa_p-4}} + K_3\frac{h^{2r_c}}{r_c^{2\kappa_c-4}} + K_4\Delta t\Big). \tag{39}$$

Equations (36) and (39) imply that $(\boldsymbol{\psi},\boldsymbol{\phi})$ belongs to $\mathcal{W}$. □

From Lemma 4 and Theorem 1, we have that the set $\mathcal{F}(\mathcal{W})$ is bounded. Using similar techniques as in Lemma 5 and Theorem 1 it can be shown that $\mathcal{F}$ is continuous. Since we are in finite dimension, this means that the operator $\mathcal{F}$ is compact. Therefore, by Schauder's fixed point theorem there is a solution $(\boldsymbol{\psi}, \boldsymbol{\phi}) \in \mathcal{W}$ such that

$$(\boldsymbol{\psi}, \boldsymbol{\phi}) = \mathcal{F}(\boldsymbol{\psi}, \boldsymbol{\phi}).$$

This fixed point solution is the DG solution to (20)-(22). Using the definition of the space $\mathcal{W}$, the approximation results (26), (27) and Lemma 4, we obtain the following *a priori* error estimates.

Theorem 2. *Let $(\boldsymbol{P}, \boldsymbol{C})$ be a solution to (20)-(22). Assume that the solution (p, c) to (1)-(6) belongs to $L^\infty(0,T; H^{\kappa_p}(\Omega)) \times L^\infty(0,T; H^{\kappa_c}(\Omega))$. Assume that the penalty parameters satisfy:*

$$\sigma_p \geq \sigma_p^*, \quad \sigma_p^* = (1-\varepsilon)^2 \frac{48(\overline{\mu})^2 \overline{k} \widetilde{M} M_t^2}{7\underline{\mu}}$$

$$\sigma_c \geq \sigma_c^*, \quad \sigma_c^* = \max\Big((1-\varepsilon)^2 \frac{21(\overline{d})^2 \widetilde{M} M_t^2}{4\underline{d}}, \frac{21(\overline{d})^2 \widetilde{M} M_t^2}{\underline{d}}\Big).$$

Then, there exists a constant M independent of $h, r_p, r_c, \Delta t$ such that for all $i \geq 1$

$$\big\|C^i - c^i\big\|_{0,\Omega} + (\Delta t \sum_{j=1}^{i} |||C^j - c^j|||_1^2)^{1/2} + |||P^i - p^i|||_1 \leq M_1\big(\frac{h^{r_p}}{r_p^{\kappa_p - 2}} + \frac{h^{r_c}}{r_c^{\kappa_c - 2}} + \Delta t\big).$$

More technical details of the convergence analysis presented above, can be found in [7].

5. Extensions and Concluding Remarks

We studied the application of primal discontinuous Galerkin methods, namely NIPG, IIPG, SIPG, and backward Euler discretization to solve the miscible displacement problem. We gave explicit expressions of the limiting values of the penalty parameters above which the method is stable and convergent. The methods, presented above can be modified slightly to consider several other boundary conditions. The convergence analysis developed in Section 4 is independent of the choice of the boundary conditions and it can be applied in the same way as above to show the stability and convergence of the scheme introduced below. For instance, we may have

$$u \cdot n = u_{\text{dir}} \quad \forall (x,t) \in \partial\Omega \times \bar{J}, \tag{40}$$

$$c = c_{\text{dir}} \quad \text{on} \quad \Gamma_{\text{D}} \times \bar{J}, \tag{41}$$

$$D(u)\nabla c \cdot n = 0, \quad \Gamma_{\text{N}} \times \bar{J}. \tag{42}$$

If (6) and (40) hold, the scheme becomes:
Pressure Equation: $\forall 0 \leq i \leq N-1$,

$$\forall z \in \mathcal{D}_{r_p}(\mathcal{E}_h), \quad \sum_{E \in \mathcal{E}_h} \int_E \frac{1}{\mu(C^{i+1})} K \nabla P^{i+1} \cdot \nabla z + \sigma_p \sum_{e \in \Gamma_h} \frac{g(r_p)}{|e|} \int_e [P^{i+1}][z]$$

$$- \sum_{e \in \Gamma_h} \int_e \{\frac{1}{\mu(C^{i+1})} K \nabla P^{i+1} \cdot n_e\}[z] + \varepsilon \sum_{e \in \Gamma_h} \int_e \{\frac{1}{\mu(C^{i+1})} K \nabla z \cdot n_e\}[P^{i+1}] = \int_\Omega f_1 z + \int_{\partial\Omega} u_{\text{dir}} z.$$

Concentration Equation: $\forall 0 \le i \le N-1$,

$$\forall v \in \mathcal{D}_{r_c}(\mathcal{E}_h),\ \int_\Omega \frac{\varphi}{\Delta t}(C^{i+1}-C^i)v + \sum_{E\in\mathcal{E}_h}\int_E \frac{C^{i+1}}{\mu(C^{i+1})}K\nabla P^{i+1}\cdot\nabla v$$

$$+\sum_{E\in\mathcal{E}_h}\int_E D(U^{i+1})\nabla C^{i+1}\cdot\nabla v - \sum_{e\in\Gamma_h}\int_e \{\frac{C^{i+1}}{\mu(C^{i+1})}K\nabla P^{i+1}\cdot n_e\}[v]$$

$$-\sum_{e\in\Gamma_h\cup\partial\Omega}\int_e \{D(U^{i+1})\nabla C^{i+1}\cdot n_e\}[v] + \varepsilon\sum_{e\in\Gamma_h}\int_e \{\frac{C^{i+1}}{\mu(C^{i+1})}K\nabla v\cdot n_e\}[P^{i+1}]$$

$$+\varepsilon\sum_{e\in\Gamma_h}\int_e \{D(U^{i+1})\nabla v\cdot n_e\}[C^{i+1}] + \sigma_c\sum_{e\in\Gamma_h\cup\partial\Omega}\frac{g(r_c)}{|e|}\int_e [C^{i+1}][v]$$

$$= \sigma_c\sum_{e\in\partial\Omega}\frac{g(r_c)}{|e|}\int_e c_{\mathrm{dir}}v + \int_\Omega f_2 v + \sum_{e\in\partial\Omega}\int_e c_{\mathrm{dir}}u_{\mathrm{dir}}v,$$

If (40), (41), (42) hold, the concentration equation becomes:
Concentration Equation: $\forall 0 \le i \le N-1$,

$$\forall v \in \mathcal{D}_{r_c}(\mathcal{E}_h),\quad \int_\Omega \frac{\varphi}{\Delta t}(C^{i+1}-C^i)v + \sum_{E\in\mathcal{E}_h}\int_E \frac{C^{i+1}}{\mu(C^{i+1})}K\nabla P^{i+1}\cdot\nabla v$$

$$+\sum_{E\in\mathcal{E}_h}\int_E D(U^{i+1})\nabla C^{i+1}\cdot\nabla v - \sum_{e\in\Gamma_h}\int_e \{\frac{C^{i+1}}{\mu(C^{i+1})}K\nabla P^{i+1}\cdot n_e\}[v] - \sum_{e\in\Gamma_{\mathrm{N}}}\int_e C^{i+1}u_{\mathrm{dir}}v$$

$$-\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}}\int_e \{D(U^{i+1})\nabla C^{i+1}\cdot n_e\}[v] + \varepsilon\sum_{e\in\Gamma_h}\int_e \{\frac{C^{i+1}}{\mu(C^{i+1})}K\nabla v\cdot n_e\}[P^{i+1}]$$

$$+\varepsilon\sum_{e\in\Gamma_h}\int_e \{D(U^{i+1})\nabla v\cdot n_e\}[C^{i+1}] + \sigma_c\sum_{e\in\Gamma_h\cup\Gamma_{\mathrm{D}}\cup\Gamma_{\mathrm{N}}}\frac{g(r_c)}{|e|}\int_e [C^{i+1}][v] =$$

$$+\sigma_c\sum_{e\in\Gamma_{\mathrm{D}}}\frac{g(r_c)}{|e|}\int_e c_{\mathrm{dir}}v + \int_\Omega f_2 v + \sum_{e\in\Gamma_{\mathrm{N}}}\int_e c_{\mathrm{dir}}u_{\mathrm{dir}}v,$$

References

[1] S. Agmon. *Lectures on Elliptic Boundary Value Problems*. Van Nostrand, Princeton, NJ, 1965.

[2] I. Babuška and M. Suri. The $h-p$ version of the finite element method with quasiuniform meshes. *RAIRO Mathematical Modeling and Numerical Analysis*, 21:199–238, 1987.

[3] C.N.Dawson, T.F.Russell, and M.F.Wheeler. Some improved error estimates for the modified method of characteristics. *SIAM J.Numer. Anal.*, 26(6):1487–1512, 1989.

[4] C. Dawson, S. Sun, and M.F. Wheeler. Compatible algorithms for coupled flow and transport. *Comput. Meth. Appl. Mech. Engng*, 193:2565–2580, 2004.

[5] Y. Epshteyn and B. Rivière. On the solution of incompressible two-phase flow by a p-version discontinuous Galerkin method. *Communications in Numerical Methods in Engineering*, 22:741–751, 2006.

[6] Y. Epshteyn and B. Rivière. Fully implicit discontinuous finite element methods for two-phase flow. *Applied Numerical Mathematics*, 57:383–401, 2007.

[7] Y. Epshteyn and B. Rivière. High order fully discrete discontinuous Galerkin methods for miscible displacement. *Center for Nonlinear Analysis (CNA) preprint*, 2007.

[8] R.E. Ewing and M.F. Wheeler. Galerkin methods for miscible displacement problems in porous media. *SIAM J. Numer. Anal.*, 17(3):351–365, June 1980.

[9] D.H. Griffel, editor. *Applied Functional Analysis*. Dover, 2002.

[10] H. Hoteit and A. Firoozabadi. Compositional modeling by the combined discontinuous Galerkin and mixed methods. *SPE Journal*, pages 19–34, March 2006.

[11] J.Douglas, M.F.Wheeler, B.L.Darlow, and R.P.Kendall. Self-adaptive finite element simulation of miscible displacement in porous media. *Computer methods in applied mechanics and engineering*, 47:131–159, 1984.

[12] J.Douglas, R.E.Ewing, and M.F.Wheeler. The approximation of the pressure by a mixed method in the simulation of miscible displacement. *R.A.I.R.O. Numerical Analysis*, 17(1):17–33, 1983.

[13] J.Douglas, R.E.Ewing, and M.F.Wheeler. A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media. *R.A.I.R.O. Numerical Analysis*, 17(3):249–265, 1983.

[14] E.J. Kubatko, J.J. Westerink, and C. Dawson. hp discontinous Galerkin methods for advection-dominated problems in shallow water. *Computer Methods in Applied Mechanics and Engineering*, 196:437–451, 2006.

[15] C. Ortner and E. Süli. Discontinuous Galerkin finite element approximation of nonlinear second-order elliptic and hyperbolic systems. *SIAM Journal on Numerical Analysis*, 45(4):1370–1397, 2007.

[16] R.E.Ewing, T.F.Russell, and M.F.Wheeler. Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics. *Comput.Methods Appl.Mech.Engrg*, 47:73–92, 1984.

[17] B. Rivière and M.F. Wheeler. Discontinuous Galerkin methods for flow and transport problems in porous media. *Communications in Numerical Methods in Engineering*, 18:63–68, 2002.

[18] B. Rivière, M.F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Computational Geosciences*, 3:337–360, April 1999.

[19] B. Rivière, M.F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM Journal on Numerical Analysis*, 39(3):902–931, 2001.

[20] S. Sun, B. Rivière, and M. Wheeler. A combined mixed finite element and discontinuous Galerkin method for miscible displacement problem in porous media. In *Recent Progress in Computational and Applied PDEs*, pages 323–348. Kluwer/Plenum, 2002.

[21] S. Sun and M.F. Wheeler. Discontinuous Galerkin methods for coupled flow and reactive transport problems. *Applied Numerical Mathematics*, 52:273–298, 2005.

[22] S. Sun and M.F. Wheeler. Analysis of discontinuous galerkin methods for multicomponent reactive transport problems. *Computers and Mathematics with Applications*, 52:637–650, 2006.

[23] T.F.Russell. Time-stepping along characteristics with incomplete iteration for a Galerkin approximation of miscible displacement in porous media. *SIAM J.Numer.Anal.*, 22:970–1013, 1985.

[24] T.Russell, M.F.Wheeler, and C.Chiang. Large-scale simulation of miscible displacement by mixed and characteristics finite element methods. *Mathematical and computational methods in seismic exploration and reservoir modeling*, pages 85–107, 1985.

[25] T. Warburton and J.S. Hesthaven. On the constants in hp-finite element trace inverse inequalities. *Comput. Methods Appl. Mech. Engrg.*, 192:2765–2773, 2003.

[26] M.F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM Journal on Numerical Analysis*, 15(1):152–161, 1978.

[27] M.F. Wheeler and B.L.Darlow. Interior penalty Galerkin procedures for miscible displacement problems in porous media. *Computational Methods in Nonlinear Mechanics*, pages 485–506, 1980.

Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA
E-mail: rina10@andrew.cmu.edu
URL: http://www.math.pitt.edu/∼yee1

Department of Mathematics, University of Pittsburgh, 301 Thackeray, Pittsburgh, PA 15260, USA
E-mail: riviere@math.pitt.edu
URL: http://www.math.pitt.edu/∼riviere

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 64–84

WAVELETS, A NUMERICAL TOOL FOR MULTISCALE PHENOMENA: FROM TWO DIMENSIONAL TURBULENCE TO ATMOSPHERIC DATA ANALYSIS.

PATRICK FISCHER AND KA-KIT TUNG

Abstract. Multiresolution methods, such as the wavelet decompositions, are increasingly used in physical applications where multiscale phenomena occur. We present in this paper two applications illustrating two different aspects of the wavelet theory.
In the first part of this paper, we recall the bases of the wavelets theory. We describe how to use the continuous wavelet decomposition for analyzing multifractal patterns. We also summarize some results about orthogonal wavelets and wavelet packets decompositions.
In the second part, we show that the wavelet packet filtering can be successfully used for analyzing two-dimensional turbulent flows. This technique allows the separation of two structures: the solid rotation part of the vortices and the remaining mainly composed of vorticity filaments. These two structures are multiscale and cannot be obtained through usual filtering methods like Fourier decompositions. The first structures are responsible for the inverse transfer of energy while the second ones are responsible for the forward transfer of enstrophy. This decomposition is performed on numerical simulations of a two dimensional channel in which an array of cylinders perturb the flow.
In the third part, we use a wavelet-based multifractal approach to describe qualitatively and quantitatively the complex temporal patterns of atmospheric data. Time series of geopotential height are used in this study. The results obtained for the stratosphere and the troposphere show that the series display two different multifractal behaviors. For large time scales (several years), the main Hölder exponent for the stratosphere and the troposphere data are negative indicating the absence of correlation. For short time scales (from few days to one year), the stratopshere series present some correlations with Hölder exponents larger than 0.5, whereas the troposhere data are much less correlated.

Key Words. Wavelets, two dimensional turbulence, multifractal analysis, atmospheric data

1. Review on wavelets

The one dimensional wavelet theory is reviewed in this part. The generalization to higher dimension is relatively easy and is based on tensor products of one dimensional basis functions. The two dimensional wavelet theory is recalled here in the wavelet packets framework only. We present here a summary of the theory, and a more complete description can be found in [12, 26].
Any time series, which can be seen as a one dimensional mathematical function, can

2000 *Mathematics Subject Classification.* 65T60, 76F65, 28A80.

be represented by a sum of fundamental or simple functions called basis functions. The most famous example, the Fourier series,

$$s(t) = \sum_{k=-\infty}^{+\infty} c_k e^{ikt} \tag{1}$$

is valid for any 2π-periodic function sufficiently smooth. Each basis function, e^{ikt} is indexed by a parameter k which is related to a frequency. In (1), $s(t)$ is written as a superposition of harmonic modes with frequencies k. The coefficients c_n are given by the integral

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} s(t) e^{-ikt} dt \tag{2}$$

Each coefficient c_k can be viewed as the average harmonic content of $s(t)$ at frequency k. Thus the Fourier decomposition gives a frequency representation of any signal. The computation of c_k is called the decomposition of s and the series on the right hand side of (1) is called the reconstruction of s.
Although this decomposition leads to good results in many cases, some disadvantages are inherent to the method. One of them is the fact that all the information concerning the time variation of the signal is completely lost in the Fourier description. For instance, a discontinuity or a localised high variation of the frequency will not be well described by the Fourier representation. The underlying reason lies in the nature of complex exponential functions used as basis functions. They all cover the entire real line, and differ only with respect to frequency. They are not suitable for representing the behaviour of a discontinuous function or a signal with high localised oscillations.
Like the complex exponential functions of the Fourier decomposition, wavelets can be used as basis functions for the representation of a signal. But, unlike the complex exponential functions, they are able to restore the positional information as well as the frequency information.

1.1. Continuous wavelets and the multifractal formalism. The wavelet-based multifractal formalism has been introduced in the nineties by Mallat [25, 26], Arneodo [2, 3, 4], Bacry [5] and Muzy [28]. A wavelet transform can focus on localized signal structures with a zooming procedure that progressively reduces the scale parameter. Singularities and irregular structures often correspond to essential information in a signal. The local signal regularity can be described by the decay of the wavelet transform amplitude across scales. Singularities can be detected by following the wavelet transform local maxima at fine scales.

The wavelet transform is a convolution product of a data sequence with the compressed (or dilated) and translated version of a basis function ψ called the wavelet mother. The scaling and translation are performed by two parameters: the scale parameter a dilates or compresses the mother wavelet to various resolutions and the translation parameter b moves the wavelet all along the sequence:

$$WT_s(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} s(t) \psi^* \left(\frac{t-b}{a} \right) dt, \ a \in \mathbb{R}^{+*}, b \in \mathbb{R}. \tag{3}$$

This definition of the wavelet transform leads to an invariant L^2 measure, and thus conserves the energy ($\|s\|_2 = \|WT_s\|_2$). A different normalization could be used leading to a different invariant.

The strength of a singularity of a function is usually defined by an exponent called Hölder exponent. The Hölder exponent $h(t_0)$ of a function s at the point t_0 is defined as the largest exponent such that there exists a polynomial $P_n(t)$ of order n satisfying:

$$|s(t) - P_n(t - t_0)| \leq C|t - t_0|^{h(t_0)}, \tag{4}$$

for t in a neighborhood of t_0. The order n of the polynomial P_n has to be as large as possible in (4). The polynomial P_n can be the Taylor expansion of s around t_0. If $n < h(t_0) < n + 1$ then s is C^n but not C^{n+1}. The exponent h evaluates the regularity of s at the point t_0. The higher the exponent h, the more regular the function s. It can be interpreted as a local measure of 'burstiness' in the time-series at time t_0. A wavelet transform can estimate this exponent by ignoring the polynomial P_n. A transcient structure or 'burst' is generally wavelet-transformed to a superposition of wavelets with the same centre of mass and wide range of frequencies.
In order to evaluate the Hölder exponent, we have to choose a wavelet mother with $m > h$ vanishing moments:

$$\int_{-\infty}^{\infty} t^k \psi(t)\, dt, \tag{5}$$

for $0 \leq k < m$. A wavelet with m vanishing moments is orthogonal to polynomials of degree $m - 1$. Since $h < m$, the polynomial P_n has a degree n at most equal to $m - 1$ and we can then show that:

$$\int_{-\infty}^{+\infty} P_n(t - t_0)\psi^* \left(\frac{t - b}{a}\right) dt \, = 0. \tag{6}$$

Let us assume that the function s can be written as a Taylor expansion around t_0:

$$s(t) = P_n(t - t_0) + C|t - t_0|^{h(t_0)} \tag{7}$$

We then obtain for its wavelet transform at t_0:

$$WT_s(t_0, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} C|t - t_0|^{h(t_0)} \psi^* \left(\frac{t - t_0}{a}\right) dt \tag{8}$$

$$= C|a|^{h(t_0)+\frac{1}{2}} \int_{-\infty}^{+\infty} |t'|^{h(t_0)} \psi(t') dt'. \tag{9}$$

We have the following power law proportionality for the wavelet transform of the singularity of $s(t_0)$:

$$|WT_s(t_0, a)| \sim a^{h(t_0)+\frac{1}{2}} \tag{10}$$

Then, we can evaluate the exponent $h(t_0)$ from a log-log plot of the wavelet transform amplitude versus the scale a.

However, we cannot compute the regularity of a multifractal signal because its singularities are not isolated. But we can still obtain the singularity spectrum of multifractals from the wavelet transform local maxima.
These maxima are located along curves in the plane (b, a). This method, introduced by Arneodo et al. [3], requires the computation of a global partition function $Z(q, a)$. Let $\{b_i(a)\}_{i\in\mathbb{Z}}$ be the position of all maxima of $|WT_s(b, a)|$ at a fixed scale a. The partition function $Z(q, a)$ is then defined by:

$$Z(q, a) = \sum_i |WT_s(b_i, a)|^q. \tag{11}$$

We can then assess the asymptotic decay $\tau(q)$ of $Z(q,a)$ at fine scales a for each $q \in \mathbb{R}$:

$$\tau(q) = \lim_{a \to 0} \inf \frac{\log Z(q,a)}{\log a}. \tag{12}$$

This last expression can be rewritten as a power law for the partition function $Z(q,a)$:

$$Z(q,a) \sim a^{\tau(q)}. \tag{13}$$

If the exponents $\tau(q)$ define a straight line then the signal is a monofractal, otherwise the signal is called multifractal: the regularity properties of the signal are inhomogeneous, and change with location.

Finding the distribution of singularities in a multifractal signal is necessary for analyzing its properties. The so-called spectrum of singularity $D(h)$ measures the repartition of singularities having different Hölder regularity. The singularity spectrum $D(h)$ gives the proportion of Hölder h type singularities that appear in the signal. A fractal signal has only one type of singularity, and its singularity spectrum is reduced to one point. The singularity spectrum $D(h)$ for any multifractal signal can be obtained from the Legendre transform of the scaling exponent $\tau(q)$ previously defined :

$$D(h) = \min_{q \in \mathbb{R}} \left(q(h + \frac{1}{2}) - \tau(q) \right). \tag{14}$$

Let us notice that this formula is only valid for functions with a convex singularity spectrum [26]. In general, the Legendre transform gives only an upper bound of $D(h)$ [18, 19]. For a convex singularity spectrum $D(h)$, its maximum

$$D(h_0) = \max_h D(h) = -\tau(0) \tag{15}$$

is the fractal dimension of the Hölder exponent h_0.

Remark: When the maximum value of the wavelet transform modulus is very small, the formulation of the partition function given in (11) can diverge for $q < 0$. A way to avoid this problem consists in replacing the value of the wavelet transform modulus at each maximum by the supremum value along the corresponding maxima line at scales smaller than a:

$$Z(q,a) = \sum_{l \in \mathcal{L}(a)} \left(\sup_{(t,a') \in l,\, a' < a} |WT_s(t,a)| \right)^q, \tag{16}$$

where $\mathcal{L}(a)$ is the set of all maxima lines l satisfying: $l \in \mathcal{L}(a)$, if $\forall a' \leq a$, $\exists (x,a') \in l$. The properties of this modified partition function are well described in [3].

1.2. One-dimensional orthogonal wavelet bases. The theoretical construction of orthogonal wavelet families is intimately related to the notion of Multiresolution Analysis [25]. A Multiresolution Analysis is a decomposition of the Hilbert space $L^2(\mathbb{R})$ of physically admissible functions (i.e square integrable functions) into a chain of closed subspaces,

$$\ldots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \ldots$$

such that

- $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$ and $\bigcup_{j \in \mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$

- $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j-1}$
- $f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0$
- There is a function $\varphi \in V_0$, called the father wavelet, such that $\{\varphi(x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_0

Let W_j be the orthogonal complementary subspace of V_j in V_{j-1}:

$$V_j \oplus W_j = V_{j-1} \tag{17}$$

This space contains the difference in information between V_j and V_{j-1}, and allows the decomposition of $L^2(\mathbb{R})$ as a direct form:

$$L^2(\mathbb{R}) = \oplus_{j\in\mathbb{Z}} W_j \tag{18}$$

Then, there exists a function $\psi \in W_0$, called the mother wavelet, such that $\{\psi(x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of W_0. The corresponding wavelet bases are then characterized by:

$$\varphi_{j,k}(x) = 2^{-j/2}\varphi(2^{-j}x-k), \quad k,j \in \mathbb{Z}, \tag{19}$$

$$\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x-k), \quad k,j \in \mathbb{Z}. \tag{20}$$

Given an integer M, it is possible to select a mother wavelet such that:

$$\int_{\mathbb{R}} dx\, \psi(x)\, x^m = 0, \quad m = 0,\ldots,M-1\ , \tag{21}$$

which means that it has M vanishing moments and the approximation order of the wavelet transform is then also M.

Since the scaling function $\varphi(x)$, and the mother wavelet $\psi(x)$ belong to V_{-1}, they admit the following expansions:

$$\varphi(x) = \sqrt{2}\sum_{k=0}^{L-1} h_k\, \varphi(2x-k), \quad h_k = \langle \varphi, \varphi_{-1,k}\rangle\ , \tag{22}$$

$$\psi(x) = \sqrt{2}\sum_{k=0}^{L-1} g_k\, \varphi(2x-k), \quad g_k = (-1)^k h_{L-k-1}\ , \tag{23}$$

where the number L of coefficients is connected to the number M of vanishing moments and is also connected to other properties that can be imposed to $\varphi(x)$. The families $\{h_k\}$ and $\{g_k\}$ form in fact a conjugate pair of quadrature filters H and G. Functions verifying (22) or (23) have their support included in $[0,\ldots,L-1]$. Furthermore, if there exists a coarsest scale, $j = n$, and a finest one, $j = 0$, the bases can be rewritten as:

$$\varphi_{j,k}(x) = \sum_{l=0}^{L-1} h_l\, \varphi_{j-1,2k+l}(x), \quad j = 1,\ldots,n\ , \tag{24}$$

and

$$\psi_{j,k}(x) = \sum_{l=0}^{L-1} g_l\, \varphi_{j-1,2k+l}(x), \quad j = 1,\ldots,n\ . \tag{25}$$

The wavelet transform of a function $f(x)$ is then given by two sets of coefficients defined as

$$d_k^j = \int_{\mathbb{R}} dx\, f(x)\, \psi_{j,k}(x)\ , \tag{26}$$

and

$$r_k^j = \int_{\mathbb{R}} dx\, f(x)\, \varphi_{j,k}(x)\ . \tag{27}$$

Starting with an initial set of coefficients r_k^0, and using (24) and (25), coefficients d_k^j and r_k^j can be computed by means of the following recursive relations:

$$d_k^j = \sum_{l=0}^{L-1} g_l\, r_{2k+l}^{j-1}\ , \tag{28}$$

and

$$r_k^j = \sum_{l=0}^{L-1} h_l\, r_{2k+l}^{j-1}\ . \tag{29}$$

Coefficients d_k^j, and r_k^j are considered in (28) and (29) as periodic sequences with the period 2^{n-j}. The set d_k^j, is composed by coefficients corresponding to the decomposition of $f(x)$ on the basis $\psi_{j,k}$ and r_k^j may be interpreted as the set of averages at various scales.

1.3. One-dimensional wavelet packets. Let H and G be a conjugate pair of quadrature filters whose the coefficients are respectively denoted by h_j and g_j. One denotes by ψ_0 and ψ_1 the corresponding father and mother wavelets. The following sequence of functions can be defined using the filters H and G:

$$\begin{aligned} \psi_{2n}(x) &= \sqrt{2}\textstyle\sum_{j\in\mathbb{Z}} h_j \psi_n(2x-j), \\ \psi_{2n+1}(x) &= \sqrt{2}\textstyle\sum_{j\in\mathbb{Z}} g_j \psi_n(2x-j). \end{aligned} \tag{30}$$

The set of these functions $\{\psi_n\}_n$ defines the wavelet packets associated to H and G. An orthonormal wavelet packet basis of $L^2(\mathbb{R})$ is any orthonormal basis selected from among the functions $2^{s/2}\psi_n(2^s x - j)$. The selection process, the so-called Best Basis algorithm, will be described in the sequel. Each basis element is characterized by three parameters: scale s, wavenumber n and position j. A useful representation of the set of wavelet packet coefficients is that of a rectangle of dyadic blocks. For instance, if one considers a signal defined at 8 points $\{x_1, ..., x_8\}$, then the wavelet packet coefficients of this function can be summarized by Table 1.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
r_1	r_2	r_3	r_4	d_1	d_2	d_3	d_4
rr_1	rr_2	dr_1	dr_2	rd_1	rd_2	dd_1	dd_2
rrr_1	drr_1	rdr_1	ddr_1	rrd_1	drd_1	rdd_1	ddd_1

TABLE 1. Dyadic blocks of wavelet packet coefficients

Each row is obtained by the application of either filter H or G to the previous row. The application of H is denoted by r as "resuming" and the application of G by d as "differencing". For instance, the set $\{rd_1 \quad rd_2\}$ is obtained by the application of the filter H to $\{d_1 \quad d_2 \quad d_3 \quad d_4\}$, and $\{dd_1 \quad dd_2\}$ by the application of the filter G. The so called Daubechies wavelets defined in [12] with several numbers of vanishing moments have been used in the sequel.

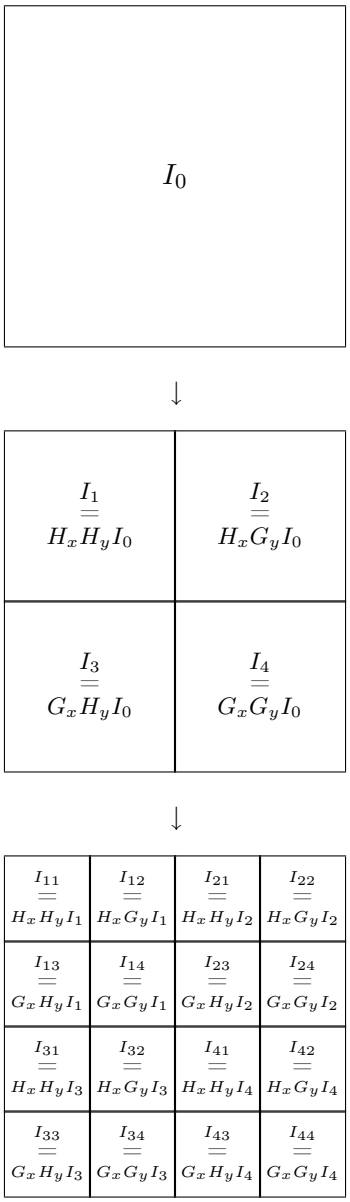


FIGURE 1. Two levels of two dimensional wavelet packets decomposition

1.4. Two-dimensional packets and the best basis algorithm. Two-dimensional wavelet packets can be obtained by tensor products $\psi_{snk}(x).\psi_{s'n'k'}(y)$ of one-dimensional basis elements. The support of these functions is exactly the cartesian product of the supports of $\psi_{snk}(x)$ and $\psi_{s'n'k'}(y)$. The same scale $s = s'$ will be used in the sequel. Subsets of such functions can be indexed by dyadic squares, with the squares corresponding to the application of one of the following filters $H \otimes H = H_x H_y$, $H \otimes G = H_x G_y$, $G \otimes H = G_x H_y$, or $G \otimes G = G_x G_y$. A graphical representation of a two-dimensional wavelet packets decomposition is given in Figure 1.

Arrays of wavelet packets constitute huge collections of basis from which one has to choose and pick. The main criterion consists in seeking a basis in which the coefficients, when rearranged into decreasing order, decrease as fast as possible. Several numerical criteria do exist and one refers to [31] for more details. The *entropy* has been chosen since it is the more often used for this type of application. For a given one-dimensional vector $u = \{u_k\}$, it is defined as:

$$E(u) = \sum_k p(k) \log(\frac{1}{p(k)}), \tag{31}$$

where $p(k) = \dfrac{|u_k|^2}{\|u\|^2}$ is the normalized energy of the k^{th} element of the vector under study. If $p(k) = 0$ then we set $p(k)\log(\frac{1}{p(k)}) = 0$. All the terms in the sum are positive. In fact, the *entropy* measures the logarithm of the number of meaningful coefficients in the original signal. The vector $p = \{p(k)\}_k$ can be seen as a discrete probability distribution function since $0 \leq p(k) \leq 1, \quad \forall k$ and $\sum_k p(k) = 1$. It can be easily shown that if only N of the values $p(k)$ are nonzero, then $E(u) \leq \log N$. Such a probability distribution function is said to be concentrated into at most N values. If $E(u)$ is small then we may conclude that u is concentrated into a few values of $p(k)$, with all other values being rare. The overabundant set of coefficients is naturally organized into a quadtree of subspaces by frequency. Every connected subtree containing the root corresponds to a different orthonormal basis. The most efficient of all the bases in the set may be found by recursive comparison: the choice algorithm will find the global minimum in $O(N)$ operations, where N is the initial degree of freedom number. In fact, the basis is chosen automatically to best represent the original data. Hence the name *best basis*. Routines in Matlab written by D. Donoho [13] and based on the algorithms designed by M.V. Wickerhauser are used for performing the packets decompositions and for searching for the best bases.

2. Application to two dimensional turbulence

While three dimensional turbulence is governed by a direct cascade of energy from the scale of injection to the small scales where the energy is dissipated, two dimensional turbulence admits two different ranges [7, 22, 23]. The first one, at large scales, is governed by an inverse energy cascade from the scale of injection to the large scales. The second one, at small scales, is governed by a cascade of enstrophy from the scale of injection to the small scales. This scenario, proposed by Kraichnan and Batchelor over 40 years ago, finds confirmation in different numerical simulations and experimental realizations. However, if the scaling laws for the different ranges have found some confirmation, the structures responsible for such transfers have not been completely identified.

Two dimensional turbulence has interested and continues to interest different scientific communities. Its relevance to atmospheric and oceanic flows at large scales has largely motivated its detailed study [24, 27, 30]. Numerical simulations have, for much longer, identified several features of 2D turbulent flows. Now, it appears that two cascades exist in a two dimensional turbulent flow. An inverse energy cascade due to the merging of same sign vortices transfers energy from the injection scale to the large scales. At scales smaller than this injection scale, an enstrophy cascade, whose origin is apparently the straining of vorticity gradients, transfers enstrophy from the large to the small scales. While the role of vortices has been identified as crucial for the dynamics of 2D flows, there has been only few if any studies of

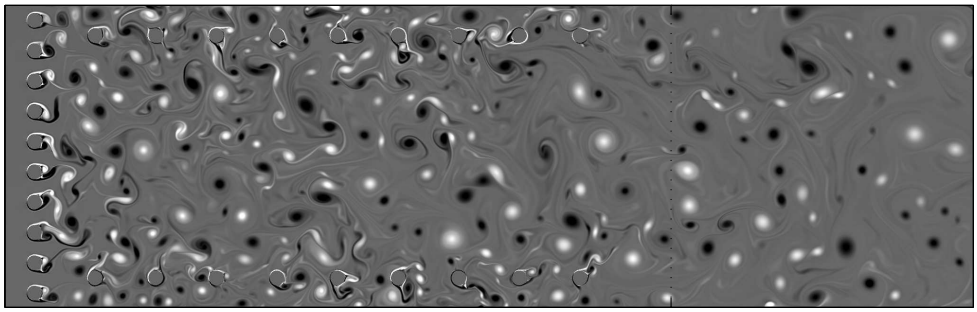


FIGURE 2. Snapshot of the vorticity field with the selected domain of analysis at the end of the channel delimited by a dotted line.

the role of flow structures on the transfers of either energy or enstrophy. This is precisely what we show here using two dimensional wavelet packets decompositions.

2.1. Numerical setup. Direct numerical simulations are used to obtain a two dimensional turbulent flow at relatively high Reynolds numbers. This flow is obtained in a channel with a length of either four or five times its width and where the turbulence is generated by arrays of cylinders. This configuration has been studied recently and the complete results have been reported in [9, 14, 15]. These simulatilons have been originally motivated by experiments carried out with soap films where grid turbulence was studied in detail [21, 20]. In order to keep a cartesian mesh, on which accurate finite differences schemes are written [11], the solid obstacles are considered as a porous medium of very weak permeability. So, instead of the classical Navier-Stokes equations, the following penalized Navier-Stokes equations [1, 10] are solved :

$$\partial_t U + (U \cdot \nabla) U \; - \; \frac{1}{Re} \, \Delta U + \frac{U}{K} + \nabla p \quad = \quad 0 \tag{32}$$

$$div U \quad = \quad 0 \tag{33}$$

where $U = (u, v)$ is the velocity, p the pressure, Re the nondimensional Reynolds number based on the unit inlet flowrate and length and K the nondimensional coefficient of permeability of the medium. The fluid and the solid media correspond to an infinite and a zero permeability coefficient respectively, $K = 10^{16}$ and $K = 10^{-8}$ are the approximate values used in the numerical simulation. The above equations are associated to no-slip boundary conditions on the walls of the channel, Poiseuille flow on the entrance section and a non reflecting boundary condition on the exit section [8]. A typical snapshot of such a simulation is presented in Figure 2 where the cylinders are apparent both near the side walls and at one distance down from the entrance. This is the flow field we analyze here using using techniques based on wavelet analysis. Contrary to standard Fourier analysis, the wavelet decomposition we use here reveals the different structures of the flow at all spatial scales. This is also different from other filtering techniques where averaging over a certain range of scales is carried out. The overall filtering process can be summarized as follows:

(1) Computation of the wavelet packets decomposition of the two components of the velocity $U = (u_1, u_2)$.

(2) Separation of the velocity fields into two subfields: the first subfield $U_s = (u_{1s}, u_{2s})$ corresponds to the wavelet packet coefficients with a modulus larger than a given threshold ϵ, and the second one $U_f = (u_{1f}, u_{2f})$ corresponds to the wavelet packet coefficients with a modulus smaller than ϵ.
(3) Construction of the corresponding vorticity fields, ω_s and ω_f. The filtered field ω_s is then essentially composed by the solid rotation part of the vortices, and the filtered field ω_f by the vorticity filaments in between that roll up in spirals inside the vortices.
(4) Computations of the physical data: energy and enstrophy spectra and fluxes.

2.2. Computation of the energy and enstrophy spectra. In this section is presented the main result concerning the analysis of the role of each filtered subfield to the two-dimensional turbulence mechanism.
The velocity decomposition $U = U_s + U_f$ obtained with the wavelet packets based filtering is orthogonal and leads to the energy spectrum decomposition

$$E(k) = E_s(k) + E_f(k), \tag{34}$$

where E_s is the energy of the solid rotation vortices and E_f is the energy of the vorticity filaments, as can be verified on Figure 3. We observe that both subfields

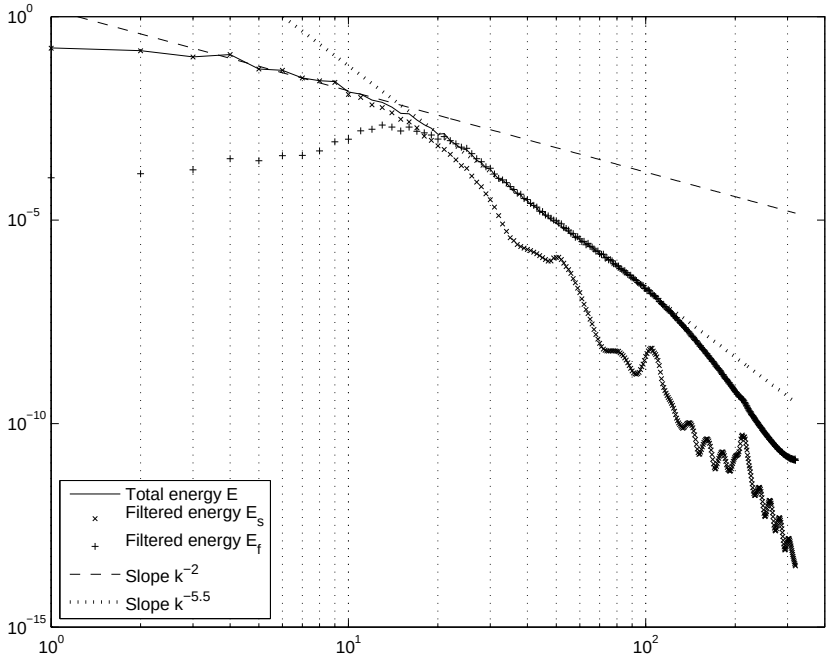


FIGURE 3. Energy spectra of the original and filtered fields obtained by a 5 scales wavelet packets decomposition ($k_{inj} \approx 20$).

are multiscale even if the E_s spectrum dominates before the injection scale and the E_f spectrum dominates after the injection scale. And the filtered energy spectra are superimposed to the global energy spectrum when they dominate. A first slope in k^{-2} and a second one in $k^{-5.5}$ on both sides of the injection scale are obtained. The first slope is not really clear as it is short but the second one is obvious.

The same decomposition of the enstrophy spectrum yields a behavior in k^0 and $k^{-3.5}$ respectively as can be observed on Figure 4. The decomposition into the

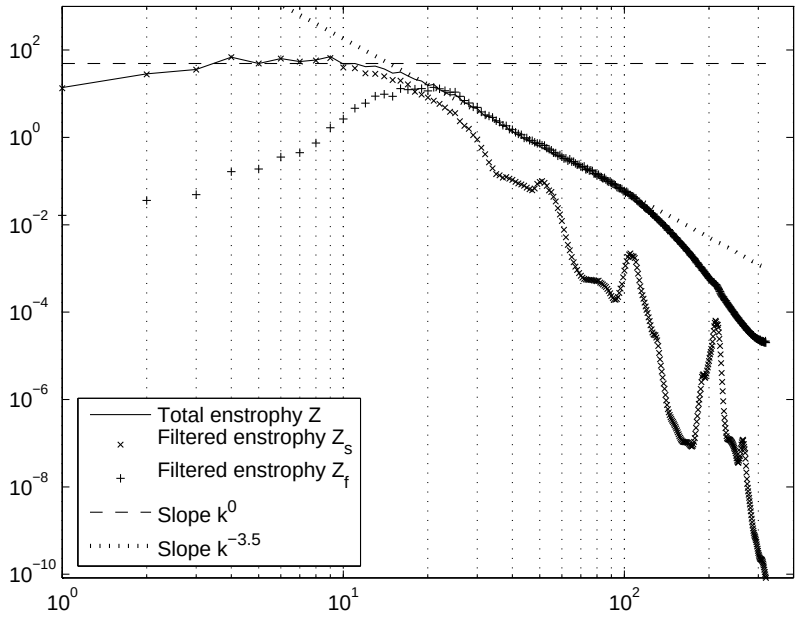
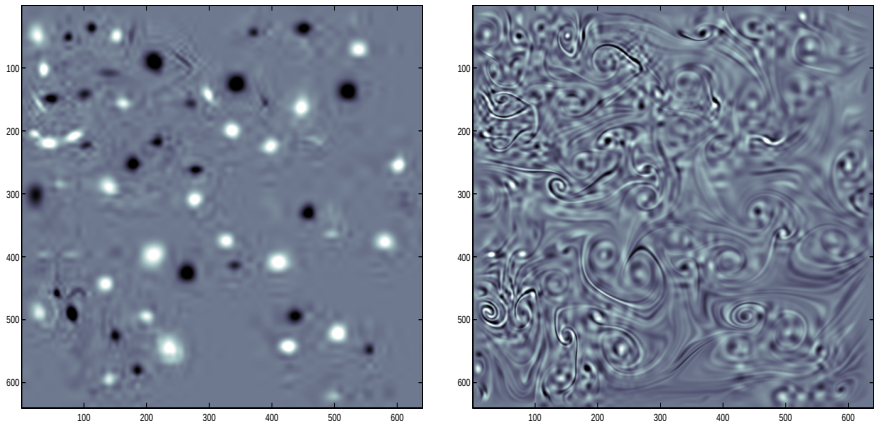


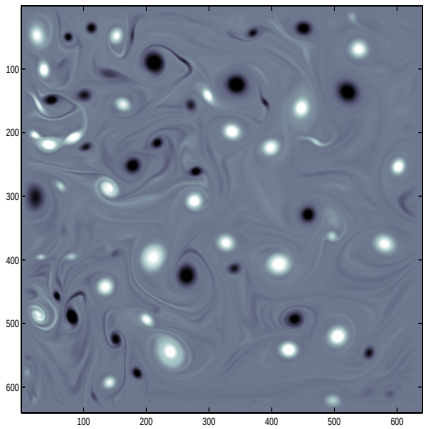
FIGURE 4. Original and filtered (WP 5 scales) enstrophy spectra ($k_{inj} \approx 20$).

two subfields obtained by the wavelet packets filtering process is given in Figure 5. The solid rotation subfield ω_s reveals all the vortices with a smooth transition and the vorticity filaments subfield ω_f shows the vorticity filaments between the vortices that end up in spirals inside the vortices. Both subfields are continuous and multiscale. The first subfield is obtained with less than 1% of the coefficients of the decomposition. It contains more than 95% of the total energy and around 70% of the enstrophy while the second one contains less than 5% of the total energy but around 30% of the enstrophy. This distribution of the enstrophy shows that unfortunately the whole flow can not be represented properly only by the first subfield. Indeed, when the vorticity filaments subfield is neglected, the global motion cannot be correct. In contrast with a Fourier based filtering, the present orthogonal filtering does not separate the scales of the flow but the type of sctructures. Here the two subfields are not seen like vortical coherent structures and background as done in previous studies but like two coherent and multiscale subfields with their own dynamics. The purpose of this paper is not the detailed study of two dimensional turbulent flows, but to show two applications of wavelet based methods. The reader particularly interested in two dimensional turbulence will find more results in [14, 15, 16].

2.3. Discussion. A careful analysis of the flows using wavelet packets filtering on sufficient levels yields relevant results one can trust. Using an adapted threshold on the wavelet coefficients allows to separate the flow into two continuous and



(a) *Solid rotations subfield* (b) *Filaments subfield*



(c) *Global vorticity field*

FIGURE 5. Wavelet packets filtering of a snapshot at the end of the channel ($k_{inj} \approx 20$).

multiscale subfields, on one hand the solid rotation of the vortices and on the other hand the vorticity filaments that connect the vortices and roll up in spirals inside the vortices. The second subfield cannot be neglected as it contains around 30% of the enstrophy and contributes for a significant part of the motion of the whole flow.

3. Multifractal analysis of atmospheric time series

Depending on the application, there are various ways of computing the wavelet transform. For the purpose of compression for instance, an orthogonal wavelet transform on dyadic scales are generally used. For the study of fractals like in this present study, continuous wavelet transforms have been found to be efficient [5]. The wavelet mother has also to be chosen according to the application. When the time series do not have any characteristic scales, or when the goal is to identify discontinuities or singularities, a real wavelet mother has to be chosen. In this

work, we use the N successive derivatives of a Gaussian function:

$$\psi(x) = \frac{d^N}{dx^N} e^{-x^2/2} \tag{35}$$

These functions are well localized in both space and frequency, and have N vanishing moments, as required for a multifractal analysis.The computations have been performed with $N = 1, 2, 4, 6, 8, 10$ but only the results for $N = 2$ are discussed in detail in this paper. The results for other values of N are very similar denoting the absence of any polynomial component. Furthermore, the case $N = 2$ is generally used for fractal analysis and corresponds to the so-called Mexican Hat function.

3.1. Data setup. We have applied the wavelet-based multifractal approach to the analysis of two sets of atmospheric data. The first set consists in the monthly averages of the NCEP Daily Global Analyzes data [29]. They correspond to times series of geopotential height from January 1948 to June 2005. A spatial average from 60°N to 90°N is performed at 17 levels, from 10 hPa down to 1000 hPa. Then the annual cycle is removed by subtracting for each month the corresponding mean in order to focus our study onto the anomalies. In such way, we will be able to detect and to describe the singularities present in the signal. Typical stratospheric and tropospheric representations are shown at 100 hPa and 700 hPa in Figures 6 and 7.

The second set of data consists in the Northern Annular Modes (NAM) at 17

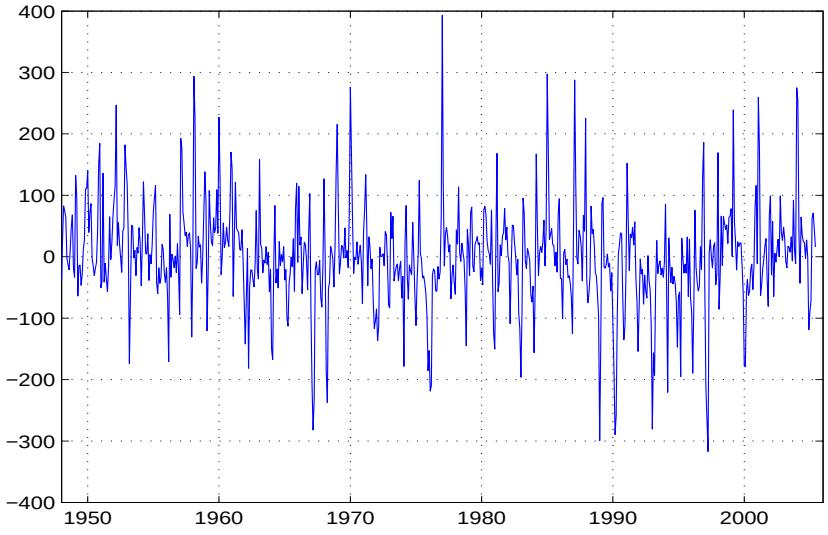


FIGURE 6. 100 hPa monthly anomalies (NCEP) from 60°N to 90°N

levels from the stratosphere down to the surface level from January 1958 to July 2006 provided by Baldwin [6]. At each pressure altitude, the annular mode is the first Empirical Orthogonal Function (EOF) of 90-day low-pass filtered geopotential anomalies north of 20°N. Daily values of the annular mode are calculated for each pressure altitude by projecting daily geopotential anomalies onto the leading EOF

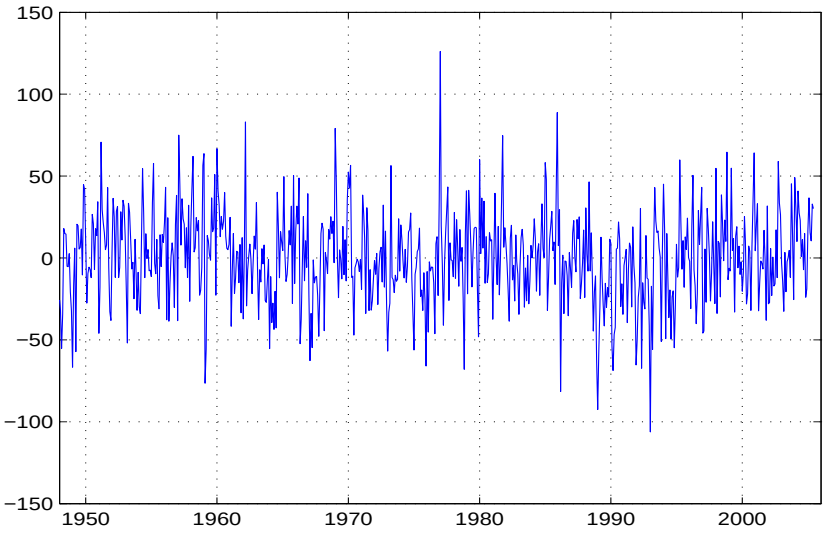


FIGURE 7. 700 hPa monthly anomalies (NCEP) from 60°N to 90°N

patterns. In the stratosphere annular mode values are a measure of the strength of the polar vortex, while the near-surface annular mode is called the Arctic Oscillation (AO), which is recognized as the North Atlantic Oscillation (NAO) over the Atlantic sector.
The results obtained with the second set of data are not given in this paper, and the reader interested in atmospheric sciences can find them in [17].

3.2. Numerical results. The wavelet decompositions obtained with the Mexican Hat function (second derivative of the Gaussian function) are given in Figures 8 and 9. The wavelet transform consists in the calculation of a resemblance index between the signal and the wavelet mother (here the Mexican Hat function). If the signal is similar to itself at different scales, then the wavelet coefficients representation will be also similar at different scales. It can be easily noticed in Figures 8 and 9 that the self-similarity generates a characteristic pattern. This representation is a good demonstration of how well the wavelet transform can reveal the fractal pattern of the atmospheric data. Based only on these representations, we cannot detect any significant difference between the stratospheric and the tropospheric signals. But we will see in the following by studying the maxima lines of the wavelet transform that these two signals have a different singularity spectrum $D(h)$.

Based on the technical reasons presented in the previous section, the partition function is computed with the formulation given in (16) for q between -20 and 20 with a step size of 0.5.
The first step in the computation of the partition function consists in the detection of the maxima lines of the wavelet transform modulus. The representation of these maxima lines, often called the "skeleton" of the wavelet transform, is given in Figure 10 for the stratospheric signal. For the computation of the partition functions,

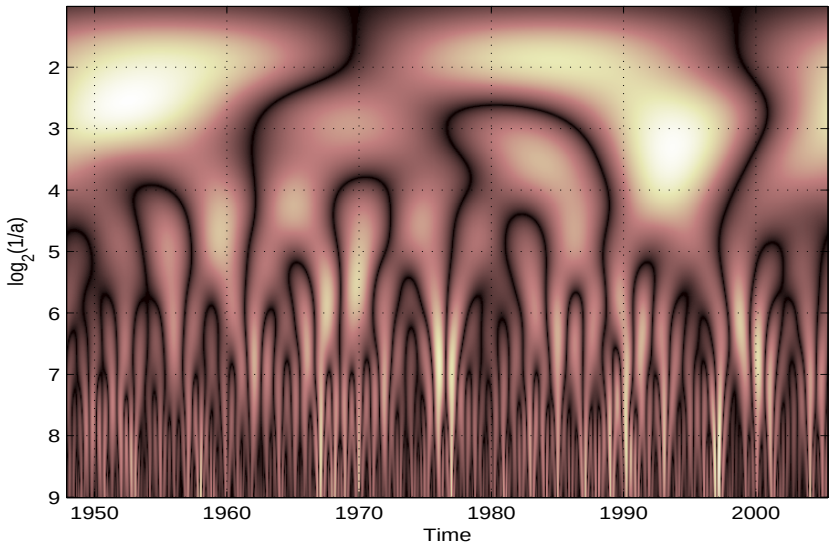


FIGURE 8. Wavelet transform modulus of the 100 hPa signal

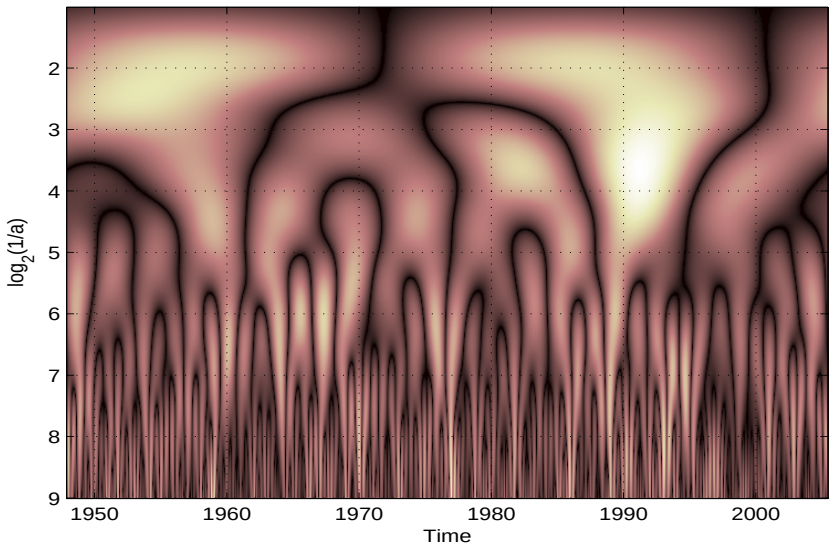


FIGURE 9. Wavelet transform modulus of the 700 hPa signal

only the maxima lines of length longer than 1 octave are kept in the summation in order to keep only the significant singularities. The two partition functions are given in Figures 11 and 12. The steps that can be observed for negative values of q are due to the use of the supremum (otherwise, the computation of $Z(q, a)$ would diverge for negative q). We can remark that the slopes for negative q are different for the stratosphere and for the troposphere. Based on this simple remark, we can

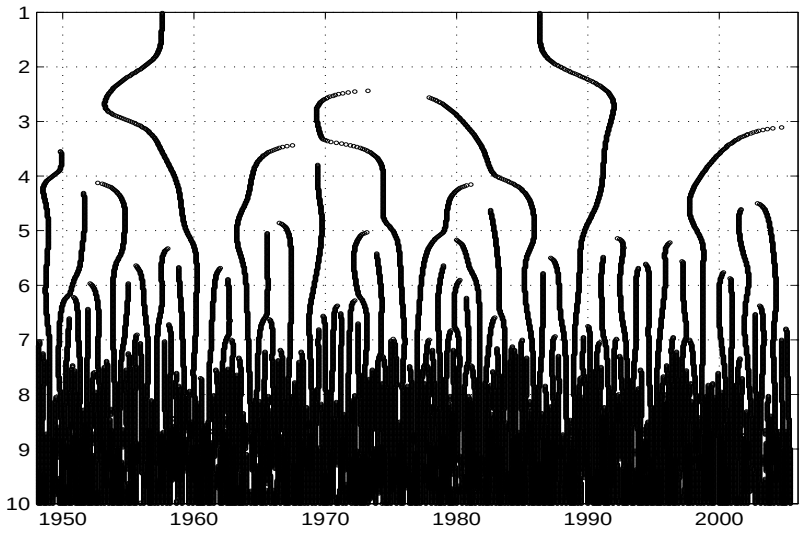


FIGURE 10. Maxima lines of the modulus of the wavelet transform of the 100 hPa signal

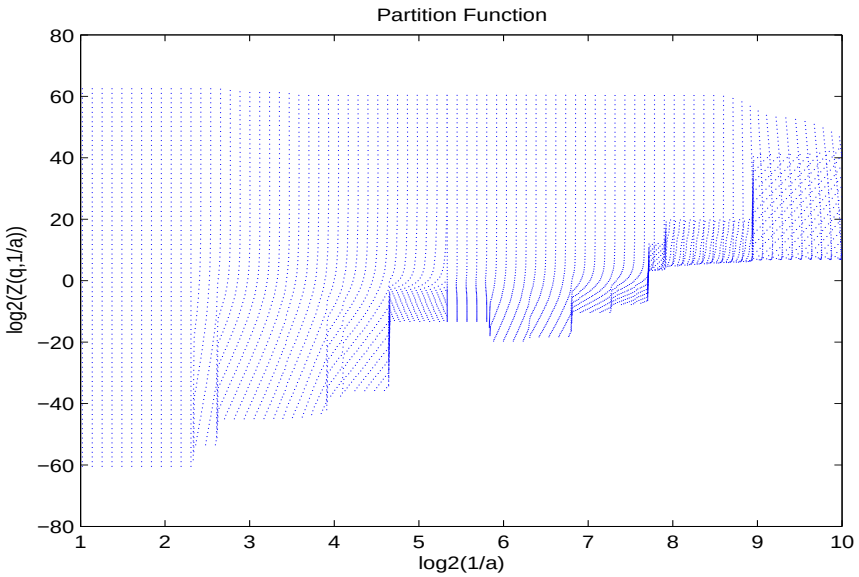


FIGURE 11. Partition function at 100 hPa

already predict that the shapes of the corresponding singularity spectra will be also different. We can expect a steeper down slope in the case of the troposphere.
The corresponding singularity spectra are given in Figure 13. The large supports of the spectra prove that the signals are multifractal. A quasi-monofractal signal spectrum would lie on very few values, and a real monofractal signal spectrum

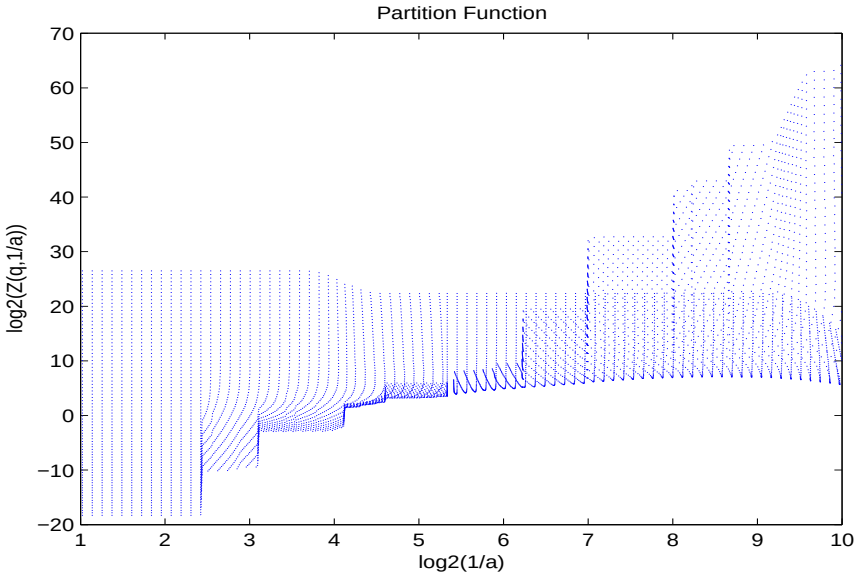


FIGURE 12. Partition function at 700 hPa

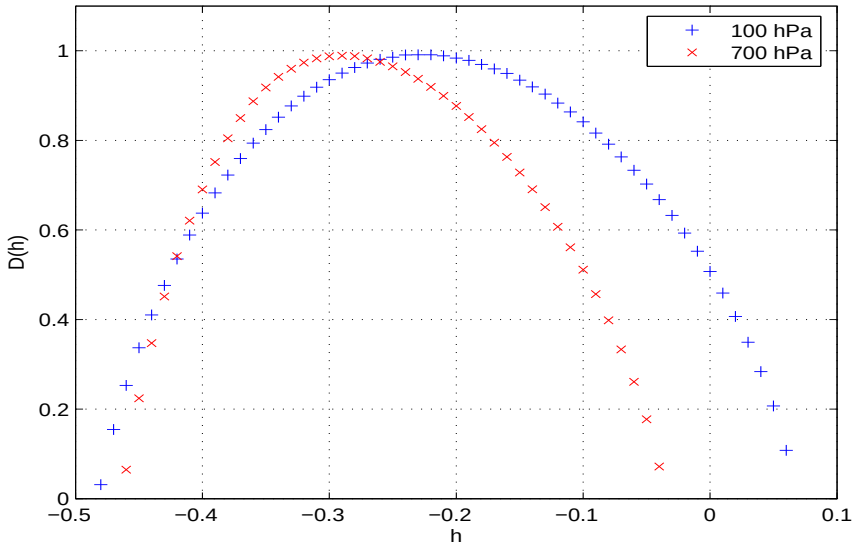


FIGURE 13. Singularity Spectra of the 100 hPa and 700 hPa signals

would reduce to only one point.
As expected, the down slope corresponding to the negative values of q is steeper for the troposphere than for the stratosphere. The maximum of the spectra is obtained around $h = -0.29$ for the stratosphere and between $h = -0.22$ and $h = -0.23$ for the troposphere. We remind here that the smaller is this value the more singular

are the singularities in the signal.

So according to this first study, we can conclude that the singularities in the tropospheric signal are more singular than the singularities in the stratospheric signal. We can verify this first conclusion by computing the value of h where the maximum of $D(h)$ is obtained for the 17 levels from 10 hPa down to 1000 hPa. The results are given in Figure 14. We can clearly detect two areas: the first one with h around -0.23 corresponds to the stratosphere and the second one with h around -0.29 corresponds to the troposphere. These results can be compared to

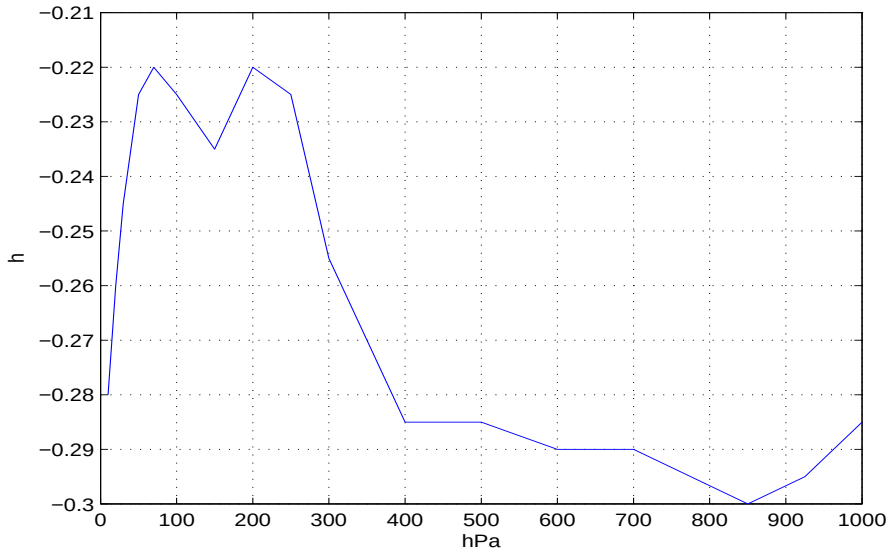


FIGURE 14. Evolution of h in function of the level

the values h obtained for artificial uncorrelated data. We perform the same computations on signals of random numbers whose elements are uniformly disctributed in the interval $(0, 1)$. The value of h found for random signals are around -0.4. So with $h \sim -0.3$ or $h \sim -0.2$, the signals corresponding to atmospheric data are close to artificial uncorrelated data at these ranges of time periods.

The whole singularity spectra can also give some information to discriminate stratospheric data from tropospheric data. We can show that their supports are also different as can be noticed from Figure 15. The stratospheric signals present broader spectra than the tropospheric signals indicating the presence of singularities over a larger spectrum.

The analysis performed on the monthly averages NCEP Data cannot give any information for periods smaller than a month. In order to get details on finer time periods, we performed the same kind of analysis on the daily NAM index. the corresponding results are given in [17].

3.3. Discussion. In this part, we have discussed some issues relating to the estimation of the multifractal nature of atmospheric data using a wavelet-based

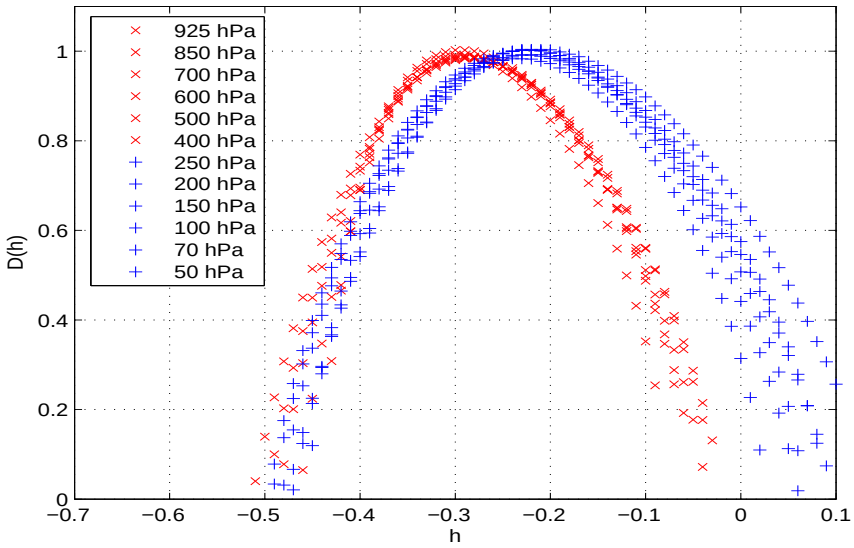


FIGURE 15. Singularity spectra for few levels in the stratosphere and in the troposphere

method. Our study reveals the clear fractal pattern of the analyzed series and their different scaling characteristics. The results obtained with daily data (not shown here) show, in the case of the stratosphere, a short-range correlation behavior that occurs for short range of time scales. In the troposphere and in the same ranges of time, we found a much weaker correlation.

4. Conclusion

Wavelets were developed independently in the fields of mathematics, quantum physics, geology and electrical engineering. They are perfect numerical tools in analyzing physical situations where the signal contains discontinuities or sharp spikes, and they are especially adapted for studying multiscale phenomena in many physical applications. We have shown in this paper a few results obtained in two different problems: two dimensional turbulence, and atmospheric data analysis. In the first application, the wavelet analysis of a two dimensional turbulent flow shows that the vorticity field can be decomposed into two orthogonal subfields. Each subfield is characterized by a distinct structure: vortices or filaments. A more detailed study [14, 15] shows that while the vortical structures are responsible for the transfer of energy upscale, the filamentary structures are responsible for the transfer of enstrophy downscale. In the second application, the continuous wavelet transform allows to enhance the multifractal patterns of the atmospheric geopotential heights. The singularity spectra of the data present different behaviors in the stratosphere and in the troposphere. The connection of the multiscaling properties of atmposheric data to the underlying physical dynamics falls beyond the scope of the present paper. However, by using a two dimensional wavelet transform, we would like to extend our research from time series to spatial patterns of atmosphere analysis.

Acknowledgments

The research was supported by the National Science Foundation, Climate Dynamics Program, under grant ATM-0332364, and the DGA (French Defense Department) under contract 06.60.018.00.470.75.01. P. Fischer would like to thank Dr. D. Casper for many fruitful conversations.

References

[1] Angot, P., Bruneau, C.H., Fabrie, P. (1999), A penalization method to take into account obstacles in incompressible viscous flows, Num. Math., **81**, 497.
[2] Arneodo, A., Grasseau, G., Holschneider, M. (1988), Wavelet transform of multifractals, Phys. Rev. Lett., **61**, 2281.
[3] Arneodo, A., Bacry, E., Muzy, J.F. (1995), The thermodynamics of fractals revisited with wavelets, Physica A, **213**, 232.
[4] Arneodo, A., Argoul, F., Bacry, E., Elezgaray, J., Muzy, J.F. (1995), Ondelettes, multifractales et turbulence, Diderot Editeur, Paris, France.
[5] Bacry, E., Muzy, J.F., Arneodo, A. (1993), Singularity spectrum of fractals signal from wavelet analysis: Exact results, J. Stat. Phys., **70**, 635-674.
[6] Baldwin, M.P., http://www.nwra.com/resumes/baldwin/nam.php
[7] Batchelor, G.K. (1969), Computation of the energy spectrum in homogeneous two-dimensional turbulence, Phys.Fluids, **12**, 233.
[8] Bruneau, C.H., Fabrie, P. (1994), Effective downstream boundary conditions for incompressible Navier-Stokes equations, Int. J. Num. Meth. Fluids, **19**, 693.
[9] Bruneau C.H., Kellay H. (2005), Coexistence of two inertial ranges in two-dimensional turbulence. Phys. Rev. E, **71**: 046305(5).
[10] Bruneau, C.H., Mortazavi, I. (2004), Passive control pf the flow around a square cylinder using porous media, Int. J. for Num. Meth. in Fluids **46**, 415.
[11] Bruneau, C.H., Saad, M. (2006), The 2D lid-driven cavity problem revisited, Computers and Fluids, **35**, 326.
[12] Daubechies, I. (1992), Ten lectures on wavelets, CBMS 61, SIAM, Philadelphia.
[13] Donoho, D. Wavelab802, Software
[14] Fischer P. (2005), Multiresolution analysis for two-dimensional turbulence. Part 1: Wavelets vs Cosine packets, a comparative study. Discrete and Continuous Dynamical Systems B, **5**, 659.
[15] Fischer P, Bruneau CH, Kellay H. (2007), Multiresolution analysis for 2D turbulence. Part 2: A physical interpretation. Discrete and Continuous Dynamical Systems B, **4**, 717.
[16] Fischer P, Bruneau CH. (2007), Spectra and filtering: a clarification, Int. J. Wavelets, Multiresolution and Information Processing, **5**, 465.
[17] Fischer P, Tung KK, Wavelet-based Multifractal Analysis of atmospheric data, Draft version.
[18] Jaffard, S. (1997) Multifractal formalism for functions Part I: Results valid for all functions, SIAM J. Math. Anal., **28**, 944.
[19] Jaffard, S. (1997) Multifractal formalism for functions Part II: Self-similar functions, SIAM J. Math. Anal., **28**, 971.
[20] Kellay, H., Wu, X.L., Goldburg, W.I. (1995), Experiments with turbulent soap films, Phys. Rev. Lett. **74**, 3975.
[21] Kellay, H., Goldburg, W.I. (2002), Two dimensional turbulence: A review of some recent experiments' Rep. Prog. Phys., **65**, 845.
[22] Kraichnan, R.H. (1967), Inertial ranges transfer in two-dimensional turbulence, Phys. Fluids, **10**, 1417.
[23] Kraichnan, R.H. (1971), Inertial-range transfer in two- and three-dimensional turbulence, J. Fluid Mech., **47**, 525.
[24] Lindborg, E. (1999), Can the atmospheric kinetic energy spectrum be explained by two-dimensional turbulence?, J. Fluid Mech. **388**, 259.
[25] Mallat, S., Zhong, S. (1991), Wavelet transform maxima and multiscale edges, in: R.M. B. et al. (Eds.), Wavelets and their Applications, Jones and Bartlett, Boston.
[26] Mallat, S. (1998) A wavelet tour of signal processing, Academic Press, New York.
[27] Morel, P., Larcheveque, M. (1974), Relative dispersion of constant-level balloons in 200mb general circulation, J. Atmos. Sci., **31**, 2189.

[28] Muzy, J.F., Bacry, E., Arneodo, A. (1991), Wavelets and multifractal formalism for singular signals: application to turbulence data, Phys. Rev. Lett., **67**,3515-3518.,
[29] NOAA-CIRES Climate Diagnostics Center in Boulder, Colorado, USA, http://www.cdc.noaa.gov
[30] Tung, K.K, Orlando, W. (2003), The k-3 and k-5/3 energy spectrum of atmospheric turbulence: Quasigeostrophic two level model simulation, J. Atmos. Sciences, **60**, 824.
[31] Wickerhauser, M. V. (1994), Adapted wavelet analysis from theory to software, A.K. Peters, Wellesley, Massachusetts.

Institut de Mathématiques de Bordeaux, Université Bordeaux 1, 33405 Talence Cedex, France
E-mail: Patrick.Fischer@math.u-bordeaux1.fr
URL: http://www.math.u-bordeaux1.fr/∼fischer

Applied Math Dept, University of Washington, Seattle, USA
E-mail: tung@amath.washington.edu
URL: http://www.amath.washington.edu/people/faculty/tung/

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 85–92

PHYSICS OF FLUID SPREADING ON ROUGH SURFACES

K. M. HAY AND M. I. DRAGILA

Abstract. n the vadose zone, fluids, which can transport contaminants, move within unsaturated rock fractures. Surface roughness has not been adequately accounted for in modeling movement of fluid in these complex systems. Many applications would benefit from an understanding of the physical mechanism behind fluid movement on rough surfaces. Presented are the results of a theoretical investigation of the effect of surface roughness on fluid spreading. The model presented classifies the regimes of spreading that occur when fluid encounters a rough surface: i) microscopic precursor film, ii) mesoscopic invasion of roughness and iii) macroscopic reaction to external forces. Theoretical diffusion-type laws based on capillarity and fluid and surface frictional resistive forces developed using different roughness shape approximations are compared to available fluid rise on roughness experiments. The theoretical diffusion-type laws are found to be the same apparent functional dependence on time; methods that account for roughness shape better explain the data as they account for more surface friction.

Key Words. roughness, wetting, capillarity

1. Introduction

The movement of fluids in unsaturated rock fractures is an involved subject, requiring an understanding of multiphase fluid dynamics and fluid interaction with soil and porous media as well as the use of complex modeling systems. However, before one can model the big picture of multiphase flow in a rock fracture system it is important to understand the basic physics that describes types of fluid movement and interaction with boundaries. In a fractured rock system, the rock surface can be porous, moist, chemically heterogeneous and rough. In this manuscript the focus will be the movement of a wetting fluid over a rough surface. Glass is commonly used to model rock when investigating characteristics of droplet movement in rock fractures. It has been observed that the speed of droplets moving down between smooth glass parallel plates is significantly different than the speed down rough glass plates and rock fractures. The physical mechanism behind fluid movement on rough surfaces is not yet well understood.

A wetting fluid is pulled into roughness by capillarity. What are the physical mechanisms that drive and resist this movement? An analytical diffusion-type law is developed that provides an explanation and a way to quantify the physical mechanisms that drive fluid invasion into roughness. The theory is based on the balance between capillary and fluid and surface frictional resistive forces. Relationships derived have the same apparent functional dependance on time as available experiments of fluid rise on roughness. The more accurate the geometry of the roughness shape, the better explain the data.

There is a large body of experimental and theoretical literature that clearly shows a rough surface affects fluid movement. Wenzel [1] observed that surface roughness caused a hydrophobic fluid to behave as if it were more hydrophobic and a hydrophilic fluid to behave as if it were more hydrophilic. Wenzel also suggested that the structure of the surface had a greater effect on the static contact angle than the chemistry. Bico *et al.* [2] suggest that a surface can be designed to tune it's wetting properties. They observe the dynamic behavior of the gas-liquid-solid interface for a hydrophilic fluid on a rough surface and derive a spreading diffusion law based on the change in energy that accompanies movement of the contact line. Cazabat and Cohen Stuart [3] explored the effects of surface roughness experimentally. They found that drops on rough surfaces spread faster than drops on smooth surfaces. While the macroscopic cap of the drop on a rough surface follows a gravity-dominated behavior, a thin fluid front rushes away from the macroscopic edge, spreading *into* the roughness by capillarity. Eventually fluid in the macroscopic drop relaxes onto the fluid film that invaded the roughness.

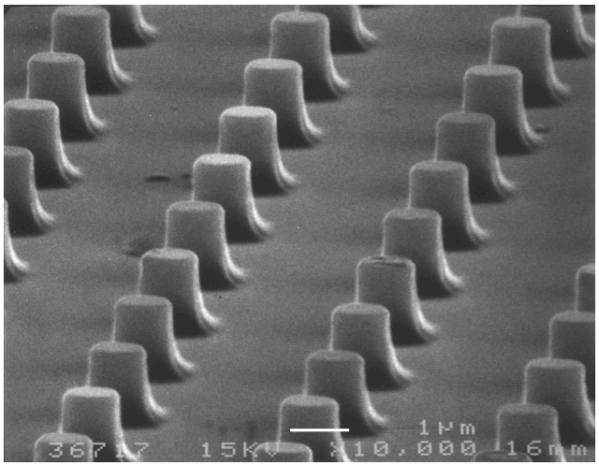


FIGURE 1. Microstructure with regular micronic cylindrical spikes used for the experiment [2].

The model for the invasion process incorporates the capillary driving mechanism suggested by experimentalists and theoreticians in this field [2], [3]. The expression derived uses an idealized geometry for the rough surface that coincides with the micropatterned surface used in experiments by Bico *et al.* [2] (Figure 1). The goal of the mathematical model is to predict the wetting behavior on a surface, given the basic surface structure and to eventually describe larger multiphase systems involving rock surfaces [4].

2. Theory

The model presented classifies three regimes of spreading: precursor film, roughness invasion and reaction to external forces (Fig. 2). It is known that a microscopic *precursor film* precedes a fluid that is in contact with a solid. Movement of the precursor film is governed by molecular diffusive transport of vacancies from the tip of the film to the edge of the macroscopic meniscus [5]. It is assumed here that the precursor film must also occur on rough surfaces and this will be considered the first regime of spreading. During the second regime of spreading on a rough

surface, the gravity-independent *mesoscopic fluid invasion*, the fluid moves into the rough texture and this regime is the focus of this study. The third regime begins when the macroscopic meniscus relaxes into its new shape or location governed by external forces, such as gravity.

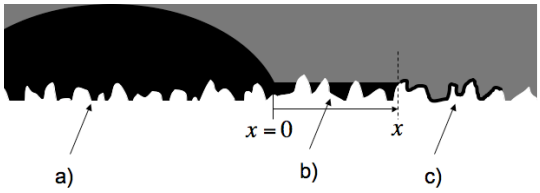


FIGURE 2. Spreading fluid drop (black) on a rough surface (white), not to scale. a) macroscopic passive drop; b) Mesoscopic fluid invasion regime (fills in the roughness); c) Precursor film (on the order of 100 Angstroms thick).

The rough surface shown in Figure 1 is idealized here as comprised of a series of small cylindrical posts lined up on an otherwise smooth surface. The fluid movement through the idealized rough surface is further simplified theoretically by modeling the surface as a series of parallel channels. This approach has been used historically in porous media by Washburn [6]. The model assumes that invasion into the rough surface is driven solely by capillary forces; gravity is ignored for this regime. Other simplifying assumptions made: the system is isothermal, solid, liquid and vapor elements are chemically homogeneous and there is no evaporation.

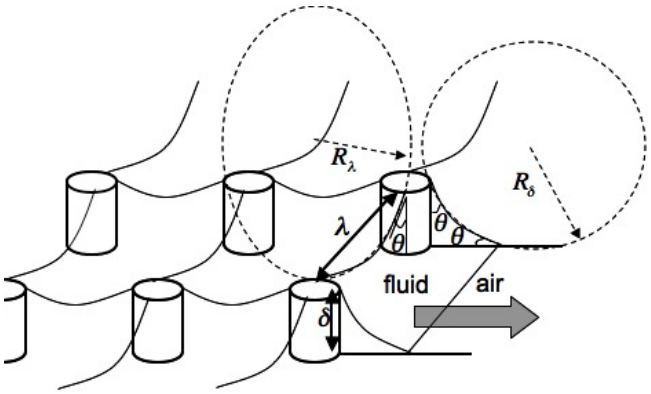


FIGURE 3. Illustration of the fluid-air interface created by the presence of roughness. All gas-fluid-solid contact angles (θ) are identical. Interface curvature creates capillarity and drives flow. Fluid movement is in the direction of the shaded arrow. Determination of the radius of curvature is governed by λ and δ.

2.1. Capillary driving mechanism. A constant capillary force exerted at the wetting front is assumed to pull fluid into roughness (invasion). The capillary force is calculated using the curvature of the fluid-gas interface. The radius of curvature is

completely described by the height of the cylinders (δ) and the separation between cylinders (λ) and the static gas-liquid-solid contact angle (θ). This force per unit area is defined by the Young-Laplace equation,

$$\Delta P_c = \gamma\Big(\frac{1}{R_\delta} + \frac{1}{R_\lambda}\Big), \tag{1}$$

where ΔP_c is the pressure difference across the fluid-air interface caused by capillarity, γ is the liquid-gas surface tension, and R_δ and R_λ are the radii of curvature. If the cylinders are normal to the smooth surface (see Figure 3) and assuming the pressure in the bulk macroscopic fluid equals the external (atmospheric) pressure then the force per unit area driving the fluid into the rough texture is [4]

$$\Delta P_c = \gamma\Big(\frac{(2\delta+\lambda)\cos\theta - \lambda\sin\theta}{\delta\lambda}\Big). \tag{2}$$

2.2. Various friction approximations. In this section, the viscous dissipation generated from fluid flow is calculated. The strength of viscous dissipation is a function of the geometry of the surface. In modeling, it is of interest to use the simplest geometry that captures the behavior. For this purpose we compare results of four geometries (of varying complexity) to the experimental results. One geometry is that of a flat surface with no roughness elements, the remainder use channel approximations.

First, a hydraulic diameter formulation is used to approximate the complex geometry of the surface (the no-slip boundary) that is in contact with the moving fluid. The hydraulic diameter method approximates a non-circular flow duct (in this case a rectangular channel) as a cylinder with an effective radius that requires knowledge of the geometry of the system, comparing the wetted area to the wetted perimeter to give a reasonable estimate of friction.

Application of the Navier-Stokes Equation for an arbitrarily shaped channel leads to a relationship between a constant pressure gradient, ΔP, across the invading fluid and the velocity of the invading fluid front, U, in the form

$$\Delta P_\mu = \frac{2Po\mu U\Delta x}{d_h^2}, \tag{3}$$

where x is the distance from the macroscopic edge of the bulk fluid to the invasion front (see Figure 2), μ is the dynamic viscosity, d_h is the hydraulic diameter and P_0 is the Poiseuille number [7]. Both d_h and P_0 are specified by the surface geometry. The pressure gradient results from the decrease in pressure in the fluid at the fluid-air interface caused by capillarity, described by Equation 2. Solving for the velocity, U, of the front edge of the invading fluid yields

$$U = \frac{\gamma d_h^2}{2Po\mu x}\Big(\frac{(2\delta+\lambda)\cos\theta - \lambda\sin\theta}{\delta\lambda}\Big). \tag{4}$$

Note that in the case of vertical imbibition, the lower boundary for the roughness driven invasion is $x_0 = \kappa^{-1}$, where κ^{-1} is the capillary length given by $(\gamma/\rho g)^{1/2}$. Even on a flat surface the wetting front will move up by a height of κ^{-1}. Integrating Equation 4 leads to a diffusion-type film invasion rate,

$$x_h = \Big[\frac{\gamma d_h^2}{Po\mu}\Big(\frac{(2\delta+\lambda)\cos\theta - \lambda\sin\theta}{\delta\lambda}\Big)\Big]^{1/2} t^{1/2} + x_0. \tag{5}$$

where t is the time it takes for the edge of the fluid to travel the distance x along the textured surface away from the macroscopic edge of the bulk fluid and the subscript h denotes the use of hydraulic diameter method to approximate the frictional resistance [4]. $x \propto t^{1/2}$ is the solution to a diffusion equation. Diffusion is

a process which describes any movement driven by an energy gradient where the energy difference is constant but the distance over which the gradient is expressed grows. Diffusion is often used to describe the process of heat transport in a fluid or mass transport in the mixing of fluids due to molecular brownian motion [7]. In this fluid invasion of roughness case, mass is transported away from the bulk fluid driven by a pressure gradient imposed by capillarity. Physically, this solution form means that the fluid movement slows as the invasion front gets further and further from the bulk fluid.

The Poiseuille number and the hydraulic diameter must be known to calculate the diffusion coefficient (the entire term in front of $t^{1/2}$ in Equation 5). The Poiseuille number, P_0, is 14.38 for a rectangle with an aspect ratio of $\alpha = \delta/\lambda = 0.48$ [8] as is the case for the experiment used here. The hydraulic diameter is [4]

$$d_h = \frac{4A}{P_w} = \frac{4\Big[\lambda\delta - \frac{\lambda^2}{4}\Big(\frac{\frac{\pi}{2}-\theta}{\cos\theta} - \tan\theta\Big)\Big]}{2\delta + \lambda}, \tag{6}$$

where the cross-sectional area of the flow, A, was approximated as a rectangle with vertical walls of height δ and width λ minus the area of a circular segment determined by the contact angle (Figure 4 e) and the wetting perimeter, $P_w = 2\delta + \lambda$. For a simple rectangle, $d_h = 4\delta\lambda/(2\delta + \lambda)$ (Figure 4 d).

If the elements are rounded, the same equation (6) can be used by substituting θ for $\theta + \theta'$ where θ' is the angle of inclination of the cylinder top (Figure 4 f) [4].

Another modeling option is to approximate the roughness as a semi circular channel. The no-slip boundary is applied along the entire semi-circular perimeter. Specifically, a *Hagen-Poiseuille half pipe flow* model gives the following function for the spreading distance,

$$x_{H-P} = \Big[\frac{\gamma\delta^2}{4\mu}\Big(\frac{(2\delta + \lambda)\cos\theta - \lambda\sin\theta}{\delta\lambda}\Big)\Big]^{1/2} t^{1/2} + x_0. \tag{7}$$

where the radius of the pipe has been approximated as the height of the obstacles, δ (see Figure 4 c), the subscript $H-P$ denotes the use of the Hagen-Poiseuille pipe flow approximation for the frictional resistance term. This approximation may be more adequate in naturally textured surfaces (e.g. rock) than for a surface such as shown in Figure 1, because natural systems are likely to have less severe vertical edges. A downfall of this method is that a pipe flow approximation will become less accurate for systems with obstruction heights not comparable to the half-width separation between obstructions. The experiment used for comparison to these theories has $\delta = 1.2\mu m$ and half-width distance, $\lambda/2 = 01.25\mu m$.

The simplest geometry is that of a *Poiseuille film flow* (Figure 4 b). The equation for the spreading distance becomes

$$x_P = \Big[\frac{2\delta^2\gamma}{3\mu}\Big(\frac{(2\delta + \lambda)\cos\theta - \lambda\sin\theta}{\delta\lambda}\Big)\Big]^{1/2} t^{1/2} + x_0, \tag{8}$$

where the subscript P denotes the use of the Poiseuille film flow approximation [4]. Approximating the flow in the roughness as a film ignores dissipation caused by the presence of the cylinders and only accounts for no-slip condition along a smooth plate (see Figure 4 b) and is thus expected to overestimate the rate of invasion.

3. Comparison to experiment

Experiments to test this concept consist of a rough microstructured surface (Figure 1) that was brought into contact with a reservoir of silicon oil. The upward

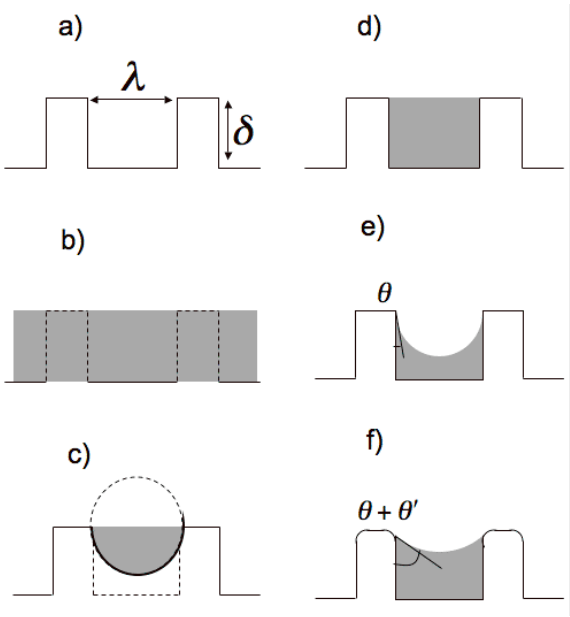


FIGURE 4. Various cross-sectional channel flow models used to estimate viscous dissipation. The solid line represents the shape of the ideal rough surface (height δ, width λ), the shaded region is the fluid. a) The rough surface, no fluid; b) Poiseuille film flow; c) Hagen-Poiseuille half-pipe flow; d) Hydraulic diameter approximation, rectangle; e) Hydraulic diameter, static contact angle off the vertical wall; f) Hydraulic diameter, static contact angle plus θ'.

rate of fluid invasion into the rough surface is not dependent on surface orientation [3], [2]. Figure 5 compares the experimental data of Bico *et al* (2001) with each of the theoretical spreading equations.

Values for the diffusion coefficients for Equations 5, 7 and 8 can be calculated using parameters from the experimental set up of Bico *et al* [2]: surface tension, $\gamma = 20.6 \times 10^{-3} N/m$; density, $\rho = 950 kg/m^3$; advancing contact angle, $\theta = 0$; dynamic viscosity, $\mu = 16 \times 10^{-3} Pa - s$; height of the cylinders, $\delta = 1.2 \times 10^{-6} m$; radius of cylinders, $R = 0.5 \times 10^{-6} m$; and distance between cylinder edges, $\lambda = 2.5 \times 10^{-6} m$ (separation between cylinder centers $= \lambda + 2R$).

Values for each theoretical diffusion coefficient correspond to the slope of the curves in Figure 5 and are listed on the figure. The data fits well to a curve that has the same functional dependence as the theories and a slope of $2.7 \times 10^{-4} m s^{-1/2}$.

3.1. Discussion. Regardless of the method used for approximating the resisting fluid friction, all the theoretical spreading equations result in diffusion laws of the form $x \propto t^{1/2}$, having the same apparent functional time dependence as the experimental data of Bico *et al*.

The half-pipe flow model is slower than the film flow model because pipe geometry provides for greater contact area although it still doesn't account for the

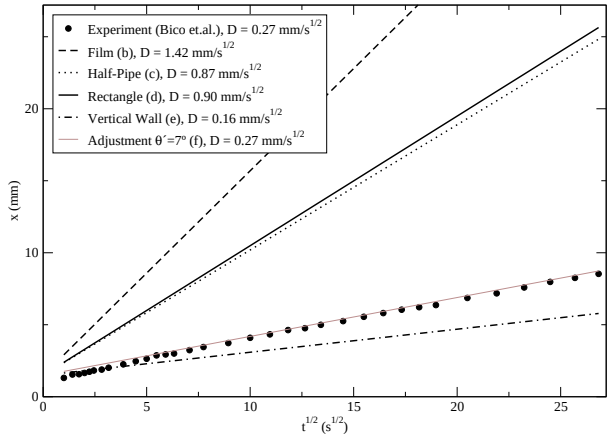


FIGURE 5. Data acquired by Bico *et al* [2]. Data is compared to various theoretical diffusion-type models, $x = Dt^{1/2} + x_0$, where D, the diffusion coefficient, equals the slope of the line. Corresponding geometry illustrations from Fig. 4 are included in the legend. The theories here have $x_0 = 1.5mm$, corresponding to the theoretical capillary length.

rectangular shape of the cylinder edges. An effective hydraulic diameter approximation provides a more realistic boundary condition than film or pipe flow and also allows for freedom in describing the approximated channel shape and fluid shape.

The hydraulic diameter approximation that uses the contact angle off the vertical wall to govern the shape predicts a diffusion coefficient that is smaller than the data. Allowing a corner angle adjustment of $\theta' = 7°$, increases the wetting area compared to wetting perimeter, and gives a close fit to the data. However, a precise match created by an arbitrary adjustment to the cylinder geometry may be fortuitous because the discrepancy between the theories and the data may be an indication of unaccounted for physical mechanisms. There remain several unaccounted for features such as an uneven advancing fluid front, small scale fluid dynamics, vertical fluid motion, and allotment for non-channelized approximation which are further discussed in Hay *et al.* [4].

4. Summary

It is the intention of this research to eventually explain the movement of the air-water interface occupying the space between rock fractures. This has applications to fluid transport through rock fractures, then on a larger scale, transport of fluid from the ground surface to groundwater, estimating the transport characteristics of contaminants. To this end, the first issue that needs to be addressed is the physical mechanism behind fluid flowing over a rough surface.

This investigation compares predictions by a theoretical model for a fluid invasion on a rough surface to experiments. The rate of spreading on a rough surface can be predicted given the surface tension, viscosity, contact angle and geometry of the surface. It is the nature of roughness in natural systems to be random and possibly fractal-like. This provides serious challenges in attempting to theoretically quantify

the wetting behavior over these surfaces. Idealizing the structure is one step closer to understanding this complex fluid movement. Results indicate that a hydraulic diameter approach, because of its flexibility in representing complex shapes may be very useful as long as the shape of the free surface is properly accounted for. Fluid movement suggested here describes the fluid invasion process that occurs when a wetting fluid encounters roughness but may not affect the overall speed of a fluid droplet between parallel plates, except by possibly changing the contact angle. This issue is being investigated further.

We thank Dr. José Bico for the use of his experimental data, Zachary Wiren for the many conceptual discussions that lead to improving the invasion theory and the NSF (Grant 0449928) for financial support.

References

[1] R. N. Wenzel, Resistance of Solid Surfaces to Wetting by Water, Ind. Eng. Chem., 28 (1936) 988-994.
[2] J. Bico, C. Tordeux, and D. Quéré, Rough Wetting, Europhys. Lett., 55 (2001) 214-220.
[3] A. M. Cazabat, and M.-A. Cohen Stuart, Dynamics of Wetting: Effects of Surface Roughness, Journal of Phys. Chem., 90 (1986) 5845-5849.
[4] K. M. Hay, M. I. Dragila, J. Liburdy, A Theoretical Model for the Wetting of a Rough Surface, To appear in Journal of Colloid and Interface Science.
[5] S. F. Burlatsky, G. Oshanin, A.-M. Cazabat, and M. Moreau, Microscopic Model of Upward Creep of an Ultrathin Wetting Film, Phys. Rev. Lett., 76 (1996) 86-89.
[6] E. W. Washburn, The Dynamics of Capillary Flow, Phys. Rev. 17 (1921) 273-283.
[7] F. M. White, Viscous Fluid Flows, McGraw-Hill, Boston, 1991.
[8] R. K. Shah, A. L. London, Advances in Heat Transfer, Academic Press, New York, NY, 1978.

Department of Physics, Oregon State University, Corvallis, Oregon

Department of Crop and Soil Sciences, Oregon State University, Corvallis, Oregon

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 93–108

A NEW PERSPECTIVE ON TEXTURE EVOLUTION

K. BARMAK, M. EMELIANENKO, D. GOLOVATY, D. KINDERLEHRER, AND S. TA'ASAN

Abstract. Modeling and analysis of texture evolution in polycrystalline materials is a major challenge in materials science. It requires understanding grain boundary or interface evolution at the network level, where topological reconfigurations (critical events) play an important role. In this paper, we investigate grain boundary evolution in a simplified one-dimensional system designed specifically to target microstructural critical event evolution. We suggest a stochastic framework that may be used to model this system and compare predictions of the model with simulations. We discuss limitations and possible extensions of this approach to higher-dimensional cases.

Key Words. Grain boundary character, Coarsening, Texture, Continuous time random walk, Boltzmann equation.

1. Introduction

Most technologically useful materials arise as polycrystalline microstructures, composed of a myriad of small crystallites, grains separated by interfaces, grain boundaries. The energetics and connectivity of the network of boundaries are implicated in many properties across all scales of use, for example, functional properties, like conductivity in microprocessor wires, and lifetime properties, like fracture toughness in structures. Engineering a microstructure to achieve a desired set of performance characteristics is a major focus in materials science. In contemporary terms, this has led to new automated data acquisition techniques, and now we are confronted with the issue of providing accurate and predictive descriptions, theories, and models. Even though this is an important and interesting subject by itself, it is also an excellent prototype for the study of multiscale phenomena.

Of course, from a multiscale viewpoint, one may aspire to begin with a molecular description of a subset of a large granular system or cellular network, and then derive a theory for its local or mesoscale behavior, and finally pass to the macroscopic state. A special advantage in our situation is that there is a well developed local thermodynamic theory based on work of Mullins [5], Herring [2], and many others, that covers normal evolution, which is the mesoscale regime. To accomplish the passage to macroscopic level, what is frequently termed upscaling in porous media networks, we need to introduce some new quantities. Indeed, it is commonly accepted that material characteristics can be traced to statistical properties of the grain boundary network. A significant advantage of the simulation platform is our ability to alter various features to assess their role or importance in a manner more flexible than nature herself permits. Historical emphasis here has been on the geometry, or more exactly, on statistics of simple geometric features of experimental

2000 *Mathematics Subject Classification.* 60K40, 82C31, 82C40.

Research supported by grants DMS 0405343 and DMR 0520425. DG acknowledges the support of DMS 0407361 and DK acknowledges the support of DMS 0305794.

and simulated polycrystalline networks, like grain area. More recently, attention has been turned to texture, the mesoscopic description of arrangement and properties of the network described in terms of both crystallography and geometry. However, the mechanisms by which the robust distributions develop from an initial population are not yet understood. As a polycrystalline configuration coarsens, facets are interchanged, some grains grow larger, and other grains disappear. Further, when triple junctions collide, new boundaries are created. We refer to these topological rearrangements as critical events. They play an important role in the evolution of distribution functions, as we explain below. In this paper, we investigate a simplified a one-dimensional system designed specifically to target critical event evolution in microstructure and its effect on texture. We use ideas from the kinetic theory of gases to study the stochastic characteristics of a one-dimensional system of grain boundaries moving under a gradient flow. We think that this model possesses some of the main features of an interacting grain boundary network in a typical polycrystalline microstructure.

In recent years, we have witnessed the introduction of automated data acquisition technologies in the materials laboratory. This has permitted the collection of statistics on a vast scale and stands to enable an important bridge between experiments and mesoscopic simulations. There are situations, for example, where it is possible to quantify the amount of alignment or misalignment sufficient to produce a corrosion resistant microstructure [1]. To rise beyond this level of anecdotal observation, the thermodynamics of the material system must be related to texture and texture related properties. Said in a different way, are there any texture related distributions which are material properties? Some geometric features of the configuration, like relative area statistics have these properties in the sense that they are robust but they are not strongly related to energetics. Recent work has provided us with a new statistic, the grain boundary character distribution, which has enormous promise in this direction. Owing to our new ability to simulate the evolution of large scale systems, we have been able to show that this statistic is robust and, in elementary cases, easily correlated to the grain boundary energy [9]—[11].

As mentioned, the regular evolution of the network of grain boundaries in two dimensions is governed by the Mullins equations of curvature-driven growth, supplemented by the Herring condition of force balance at triple junctions—a system of parabolic equations with natural boundary conditions [12]–[7]. For the higher dimensional formulation of capillary driven growth, see [8]. When applied to a single evolving $n-$sided grain with constant grain boundary energy, this mechanism leads to the Mullins-von Neumann $n-6$ rule [3]—the rate of change of the area of the grain is proportional to $n-6$, i.e.,

$$\frac{dA_n}{dt} = \gamma(n-6) \text{ where } A_n \text{ is the area of an } n\text{-sided grain,} \tag{1}$$

and $\gamma > 0$ is some material constant. MacPherson and Srolovitz [13] have given, very recently, higher dimensional generalizations of the $n-6$ rule. In particular, from (1), grains with $3, 4$ or 5 sides decrease in area. When averaged over a population of grains, equation (1) results in

$$\frac{d\bar{A}_n}{dt} = \gamma(n-6) \text{ where } \bar{A}_n \text{ is the average area of } n\text{-sided grains.} \tag{2}$$

Inspection of Fig. 1 shows that, contrary to (2), the average area of five-sided grains in a columnar aluminum structure increases several fold over the course of

an annealing experiment. Stagnation is also present in the experiment, but this is a different matter. The $n-6$-rule does not fail for the continuous changes of boundary positions, but most of the five-sided grains we observe at time $t = 2$ hours had $6, 7, 8, ...$ sides at some earlier time $t < 2$ hours. Thus in the network setting, the critical events of grain deletion and side interchange play a major role the precise mechanism of which is not yet understood.

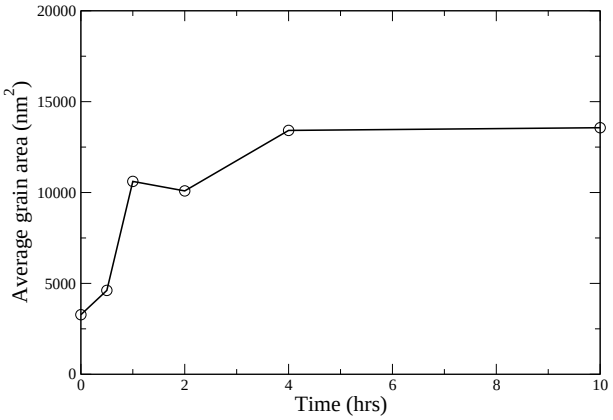


FIGURE 1. Average area of five-sided grains in an Al columnar structure.

Said differently, the grain boundary character distribution suggests that boundaries with high interfacial energies tend to shrink, while those with lower energies tend to grow. On the other hand, in a system with only geometric evolution, a grain grows or shrinks depending on its number of sides. These two situations represent extremes of behavior and in reality both effects should be taken into account. Impressive computational results have been obtained recently by [14] for a type of a birth-death model in the case of a sharp Read-Schockley type of grain boundary energy potential.

To gain an insight into the influence of critical events on the coarsening dynamics, here we model the evolution of statistical characteristics of a relatively simple, one-dimensional system of grain boundaries. The model preserves features of an interacting grain boundary network—boundaries and junctions between boundaries moving under a form of a gradient flow.

The one-dimensional model was introduced in [17] and [18] to investigate the critical events that occur during interface evolution. In particular, we used probabilistic arguments to develop a statistical model for critical events and investigated the applicability of a fractional continuous time random walk theory. Even though the fractional random walk dynamics appears to be appropriate for approximating some intermediate regimes in the evolution of the grain boundary system, the "slowing-down" coarsening effects require a more general stochastic framework. A possible approach identified in [18] is based on identifying the stochastic features of the system and formulating and solving the appropriate master equation.

In this paper we introduce an alternative framework based on a statistical mechanics approach (Section 4). To describe the evolution of statistical characteristics

of the one-dimensional system of grain boundaries, we propose two kinetic models that differ in their choice of the underlying phase space (the type and the number of state variables needed to describe an individual boundary). Both models lead to a Boltzmann-type equation for a number density of states. The numerical solution of these equations qualitatively reproduce the distributions obtained via simulation of the deterministic system of grain boundaries. Not unexpectedly, the quantitative predictions of the equation based on a larger state space are more accurate but also computationally more expensive.

2. One-dimensional model

Our principal goal is to understand whether it is possible to derive a stochastic model of grain growth by conducting numerical experiments for a large number of evolving grains, collecting the appropriate statistical data, and using this data to formulate a mathematical model governing the evolution of relevant effective characteristics. In this paper, we demonstrate the feasibility of this approach by introducing a one-dimensional system of grain boundaries represented by intervals on a number line. Note that we do not claim that such a system is physically realistic—there is no curvature-driven propagation in one dimension. Rather, our interest is in studying the dynamics of a system where interactions between grain boundaries resemble qualitatively those observed in a real polycrystalline material. We assume that each grain boundary is described by its length and a prescribed "orientation". We require that there are only nearest-neighbor interactions between the grain boundaries and that the strength of the interactions depends on values of the orientation parameter for the neighboring boundaries.

To make our system precise, fix $L > 0$ and consider the intervals $[x_i, x_{i+1}]$, $i = 0, \dots, n-1$ on the real line where $x_i \leq x_{i+1},\ i = 0, \dots, n-1$ and $x_n = x_0 + L$. The locations of the endpoints $x_i,\ i = 0, \dots, n$ may vary in time and the total length L of all intervals remains fixed. For each interval $[x_i, x_{i+1}]$, $i = 0, \dots, n-1$, choose a number α_i from the set $\{\alpha_j\}_{j=1,\dots,n}$. The intervals $[x_i, x_{i+1}]$ correspond to grain boundaries and the points x_i represent the triple junctions. The parameters $\{\alpha_i\}_{i=1,\dots,n}$ can be viewed as representing crystallographic *orientations*. The *length* of the i^{th} grain boundary is given by $l_i = x_{i+1} - x_i$. Now choose a non-negative energy density $f(\alpha)$ and define the energy

$$En(t) = \sum f(\alpha_i)(x_{i+1}(t) - x_i(t)) \tag{3}$$

Consider gradient flow dynamics characterized by the system of ordinary differential equations

$$\dot{x}_i = f(\alpha_i) - f(\alpha_{i-1}),\ i = 0, \dots, n. \tag{4}$$

The parameter α_i is prescribed for each grain boundary initially according to some random distribution and does not change during its lifetime. The *velocities* of the grain boundaries can be computed from the relation

$$v_i = \dot{x}_{i+1} - \dot{x}_i = f(\alpha_{i+1}) + f(\alpha_{i-1}) - 2f(\alpha_i). \tag{5}$$

Notice that the velocities remain constant until the moment of a *critical event* when a neighboring grain boundary collapses, at which instant a jump of the velocity occurs. Every critical event changes the statistical state of the model through its effect on the grain boundary velocities and, therefore, affects further evolution of the grains. Notice that the lengths of the individual grain boundaries vary linearly with time between the corresponding jump events with the rate that depends entirely on the corresponding grain boundary velocities.

An important feature of the thermodynamics of grain growth is that it is dissipative for the energy during normal grain growth, [12]. At critical events, the algorithm (4) is designed to enforce dissipation. To verify that (4) is dissipative, first consider a time t between two critical events. Then

$$\begin{aligned}\frac{dEn}{dt}(t) &= \sum f(\alpha_i)v_i = \sum f(\alpha_i)(f(\alpha_{i+1}) + f(\alpha_{i-1}) - 2f(\alpha_i)) \\ &\leqq 2\Big(\sum f(\alpha_i)^2\Big)^{\frac{1}{2}}\Big(\sum f(\alpha_i)^2\Big)^{\frac{1}{2}} - 2\sum f(\alpha_i)^2 = 0\end{aligned}$$

by periodicity and the Schwarz Inequality. This also corresponds to the fact that for any gradient flow dynamics

$$(\dot{x}_i)^2 = -\frac{\partial En}{\partial x_i}\dot{x}_i, \tag{6}$$

so that

$$\frac{\partial En}{\partial t} = -\sum \dot{x}_i^2 < 0. \tag{7}$$

Now suppose that the grain boundary $[x_c, x_{c+1}]$ vanishes at time $t = t_c$ and it is the only grain boundary vanishing at t_c. Then the velocity of that boundary $v_c(t) < 0,\ t < t_c$, namely,

$$\frac{1}{2}(f(\alpha_{c+1}) + f(\alpha_{c-1})) < f(\alpha_c). \tag{8}$$

and $l_c \to 0$ for $t \to t_c^-$. Now

$$En(t) > \sum_{i\neq c} f(\alpha_i)l_i,\ t < t_{crit}, \tag{9}$$

and

$$En(t_{crit}) = \lim_{t\to t_{crit}} \sum_{i\neq c} f(\alpha_i)l_i \leqq \lim_{t\to t_{crit}} En(t). \tag{10}$$

Thus the model system is dissipative.

From the materials science perspective, it is important to know both the distributions of relative lengths, as well as the grain orientations. In the most general case, we have a state space $S = \{(l, v, \alpha)\}$, where $l \in \mathbb{R}^+$, $v \in \mathbb{R}$, and $\alpha \in (a, b)$.

Our goal is to obtain (if possible) a set of equations describing time evolution of the joint probability density function (pdf) $\rho(l, v, \alpha, t)$ and, therefore, the effective dynamics of the deterministic one-dimensional grain growth model associated with (4).

3. Simulation statistics

As a first step toward a mesoscopic model we identify the set of stable statistics by simulating the one-dimensional system that evolves according to (5). The statistics of several numerical experiments for a system of 10000 grain boundaries is presented in Figures 2 and 3. Note that, unless there are coincident events, 10000 grain boundaries disappear exactly after 10000 critical events.

Figure 2 shows evolution of the relative area and the relative velocity distributions for the case of a single-well potential (the distributions are similar for other choices of f and, indeed, resemble the two dimensional statistics reported in [11]). Both statistics do not change their overall shape in the later part of the simulation, however, their spread narrows with time because fewer and fewer grain boundaries remain in the system. If the axes are scaled accordingly, we observe the stabilization of both distributions (Figure 2).

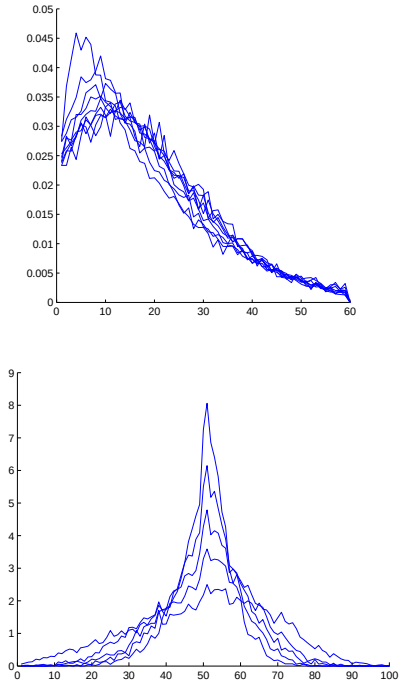


FIGURE 2. Evolution of marginal probability density functions for: (a) relative length, (b) relative velocities when $f = (x - 0.5)^2$

In the Figure 3 we present the distributions for the orientation parameters α when f has either one or two minima. The graphs clearly show that the shapes of f and orientations distribution are inversely correlated.

4. Boltzmann-type kinetic equation

4.1. Grain boundary network as a network of interacting particles. Adopting a statistical mechanics [20], [19] perspective, we will regard the system of grain boundaries as a collection of interacting "particles", where the state of each particle is determined by the parameters of the corresponding boundary. Some of these parameters—such as the length l—vary continuously with time and some—such as the orientation α and the velocity v—can change only when a grain boundary disappears during a critical event. From now on, exploiting the grain-boundary/particle analogy, we will also refer to the critical events as "collisions". Note that exactly three grain boundaries are involved in each collision—one boundary disappears while two of its immediate neighbors come in contact.

We will use the following set of conventions in order to distinguish between various types of colliding grain boundaries (Figure 4):

(1) The parameters of a grain boundary that exists prior to and is involved in a collision will carry an asterisk, i.e. (l^*, v^*, α^*).
(2) For every $i, j \in \mathbf{N}$, the subscripts $(+i)$ and $(-j)$ will be used, respectively, to label the parameters of the i-th right and the j-th left neighbor of a given grain boundary.

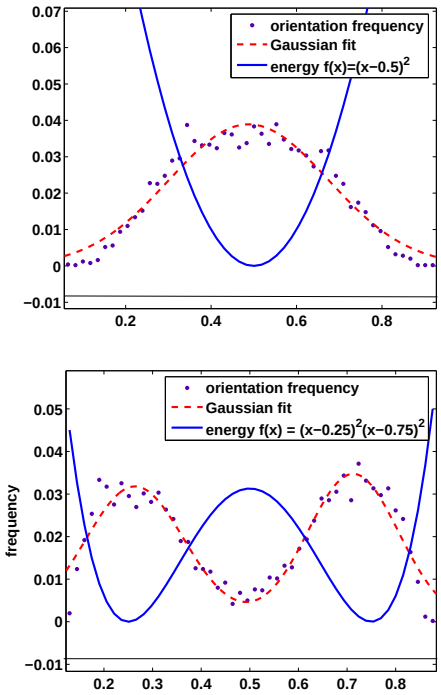


FIGURE 3. Probability density functions of the orientation parameter α for the two different choice of energy density. (a) $f(x) = (x-0.5)^2$, (b) $f(x) = (x-0.5)^2(x-1)^2(x-1.5)^2$.

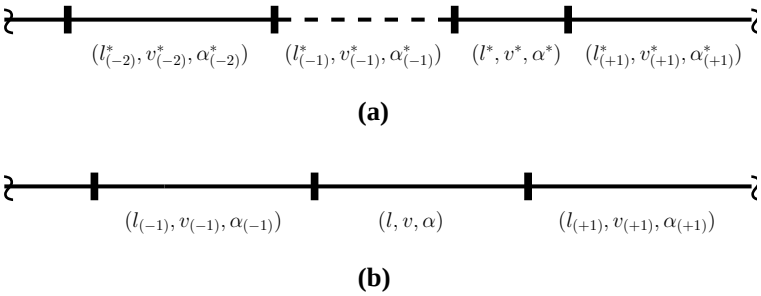


FIGURE 4. Labeling of the grain boundaries (a) before and (b) after a collision. The disappearing boundary is represented by the dashed line.

We will make extensive use of the relationship (5) which can be written as

$$v = f\left(\alpha_{(-1)}\right) + f\left(\alpha_{(+1)}\right) - 2f(\alpha), \tag{11}$$

in the new notation.

In what follows, we will consider two possible choices of a phase space for the one-dimensional grain boundary system: we will assume that a grain boundary is characterized by a set of either three (l, v, α) (model A) or four $(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)})$ (model B) variables. Observe that the lower-dimensional phase space formulation

possibly carries less information about the state of the grain boundary network, but may lead to a computationally less expensive model.

4.2. Model A.

4.2.1. Collision rules. In order to formulate the appropriate kinetic equations we need to define the collision rules that relate the parameters of the new grain boundaries that form in collisions.

Suppose that the length of the first neighbor to the left of a boundary (l^*, v^*, α^*) shrinks to zero at a time t. According to (11) the following relationships hold immediately before the collision

$$v^* = f\left(\alpha^*_{(+1)}\right) + f\left(\alpha^*_{(-1)}\right) - 2f(\alpha^*), \tag{12}$$

$$v^*_{(-1)} = f\left(\alpha^*\right) + f\left(\alpha^*_{(-2)}\right) - 2f\left(\alpha^*_{(-1)}\right). \tag{13}$$

At the time of the collision, the grain boundary $\left(0, v^*_{(-1)}, \alpha^*_{(-1)}\right)$ disappears and the boundaries (l^*, v^*, α^*) and $\left(l^*_{(-2)}, v^*_{(-2)}, \alpha^*_{(-2)}\right)$ come in contact to form the two new grain boundaries, (l, v, α) and $\left(l_{(-1)}, v_{(-1)}, \alpha_{(-1)}\right)$. Both the lengths and the orientations of the boundaries that have existed prior to the collision, transfer without change to those of new boundaries, in particular,

$$\alpha = \alpha^*,\ l = l^*,\ \alpha_{(-1)} = \alpha^*_{(-2)}. \tag{14}$$

The velocity of the boundary (l, v, α) that replaces (l^*, v^*, α^*) can be determined from (11) and is given by

$$v = f(\alpha_{(-1)}) + f(\alpha_{(+1)}) - 2f(\alpha). \tag{15}$$

From (12)-(15) we obtain the relationship

$$v = v^* + v^*_{(-1)} + f\left(\alpha^*_{(-1)}\right) - f(\alpha), \tag{16}$$

between the velocity v of the new grain boundary and the parameters of the two boundaries—$\left(0, v^*_{(-1)}, \alpha^*_{(-1)}\right)$ and (l^*, v^*, α^*)—that have collided at the time t. The first two equations in (14) and the equation (16) can be interpreted as the closed set of "collision" rules that define the grain boundary (l, v, α) in terms of parameters of its colliding "parent" boundaries $\left(0, v^*_{(-1)}, \alpha^*_{(-1)}\right)$ and (l^*, v^*, α^*). Note that this kind of collision dynamics resembles the "sticky" collisions of completely inelastic particles observed, for example, in granular gases [21].

Without loss of generality, we assume that the potential f satisfies $\min f = 0$. By (11), we have for any grain boundary (l, v, α) that

$$v + 2f(\alpha) = f(\alpha_{(-1)}) + f(\alpha_{(+1)}) \geq 0.$$

The admissible set in the phase space is given by

$$\mathcal{A} := \{(l, v, \alpha) \mid l \geq 0,\ v + 2f(\alpha) \geq 0\}. \tag{17}$$

When $\left(0, v^*_{(-1)}, \alpha^*_{(-1)}\right)$ collides from the left with (l^*, v^*, α^*) to form (l, v, α), the equations (12)-(14) provide the following constraints on the parameters of the colliding boundaries

$$\begin{cases} v^* + 2f(\alpha) - f\left(\alpha^*_{(-1)}\right) = f\left(\alpha^*_{(+1)}\right) \geq 0, \\ v^*_{(-1)} + 2f\left(\alpha^*_{(-1)}\right) - f\left(\alpha\right) = f\left(\alpha^*_{(-2)}\right) \geq 0. \end{cases} \tag{18}$$

By (15) and (18) the new boundary (l, v, α) satisfies $v+2f(\alpha) \geq 0$, thus $(l, v, \alpha) \in \mathcal{A}$.

By eliminating v^* from the first inequality in (18) and using (16), we find that $v^*_{(-1)}+2f\left(\alpha^*_{(-1)}\right) \leq v+3f(\alpha)$. Combining this inequality with the second inequality in (18) we obtain the set of constraints

$$f(\alpha) \leq v^*_{(-1)} + 2f\left(\alpha^*_{(-1)}\right) \leq v + 3f(\alpha), \tag{19}$$

on the admissible values of the parameters of a grain boundary the collision of which from the left with another grain boundary would result in formation of a grain boundary residing in the state (l, v, α).

The dynamics of a density of states in the phase space is determined by the continuous evolution of lengths between the collision events and discrete changes in velocities and orientations during these events. Next we formulate the equation that describes the evolution of the density function.

4.2.2. Evolution equation. Let $N(t)$ be the number of grain boundaries in the system at time t and set $N_0 := N(0)$. Note that the function N is non-increasing. Now suppose that $\rho(l, v, \alpha, t)$ represents the number density of states of the grain boundary system at a time t and satisfies

$$\int_{\mathcal{A}} \rho(l, v, \alpha, t)\, dl\, dv\, d\alpha = N(t), \tag{20}$$

that is $N_0^{-1}\rho(l, v, \alpha, t)dl\, dv\, d\alpha$ is a fraction of the initial number of the grain boundaries that are still present in an element $[l, l + dl] \times [v, v + dv] \times [\alpha, \alpha + d\alpha]$ of the phase space at the time t.

Since we interpret collisions as occurring between two boundaries one of which shrinks to a point at the time of the collision, we will simplify the notation by labeling the shrinking boundary as $(0, v', \alpha')$ as shown in Figure 5.

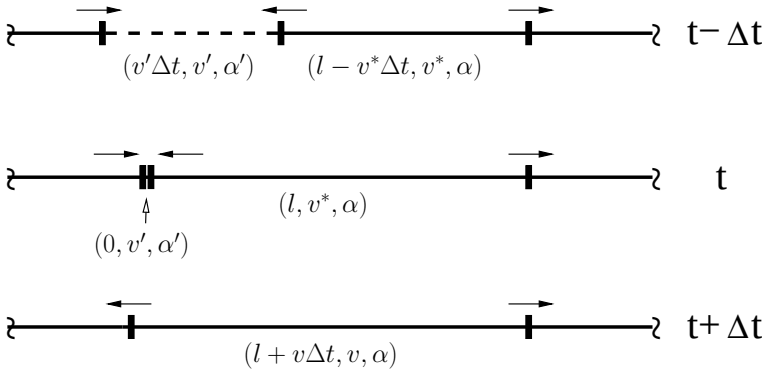


FIGURE 5. Schematic of a collision event. For each junction between the boundaries, the arrow indicates the direction of motion. The velocity $v = v^* + v' + f(\alpha') - f(\alpha)$ by (16).

Then, using (16), (19), and the fact that the grain boundaries may disappear with equal probability both to the left and to the right of a given boundary, we find that the rate per unit volume of the phase space at which the boundaries are created in $[l, l + dl] \times [v, v + dv] \times [\alpha, \alpha + d\alpha]$ is given by

$$W_+ := -\frac{1}{N(t)} \int_{A_+} v' \rho\,(0, v', \alpha', t)\, \rho\,(l, v - v' + f(\alpha) - f(\alpha'), \alpha, t)\; d\alpha' dv', \tag{21}$$

where $A_+ := \{f(\alpha) \le v' + 2f(\alpha') \le v + 3f(\alpha)\} \cap \{v' < 0\}$. Here the second restriction on the domain of integration is due to the requirement that v' must be negative to ensure that (l', v', α') is shrinking.

Now suppose that the grain boundaries $(0, v', \alpha')$ and (l, v, α) collide to form the boundary (l, w, α), where

$$w = v + v' + f(\alpha') - f(\alpha), \tag{22}$$

(cf. (16)). The collision is feasible if

$$\begin{cases} v' + 2f(\alpha') - f(\alpha) \ge 0, \\ v + 2f(\alpha) - f(\alpha') \ge 0, \end{cases} \tag{23}$$

(cf. (18)) and $(0, v', \alpha')$ must satisfy the constraints

$$v' + 2f(\alpha') \ge f(\alpha), \;\; f(\alpha') \le v + 2f(\alpha). \tag{24}$$

Then the rate per unit volume of the phase space at which the boundaries are removed from the element $[l, l + dl] \times [v, v + dv] \times [\alpha, \alpha + d\alpha]$ is given by

$$W_- := -\frac{1}{N(t)} \int_{A_-} v' \rho(0, v', \alpha', t) \rho(l, v, \alpha, t)\, d\alpha' dv', \tag{25}$$

where $A_- := \{f(\alpha) \le v' + 2f(\alpha')\} \cap \{f(\alpha') \le v + 2f(\alpha)\} \cap \{v' < 0\}$.

By taking into account the flux across the boundary of the element of the phase space (due to continuous dependence of lengths of the grain boundaries on time), we arrive at the following form of the kinetic equation

$$\frac{\partial \rho(l, v, \alpha, t)}{\partial t} + v \frac{\partial \rho(l, v, \alpha, t)}{\partial l} = W. \tag{26}$$

Here the term on the right hand side accounts for the changes in the population due to collisions

$$W := W_+ - W_- = \{gain\} - \{loss\}. \tag{27}$$

Then

$$\begin{aligned} &\frac{\partial \rho(l, v, \alpha, t)}{\partial t} + v \frac{\partial \rho(l, v, \alpha, t)}{\partial l} = \\ &-\frac{1}{N(t)} \int_{A_+} v' \rho(0, v', \alpha', t) \rho(l, v - v' + f(\alpha) - f(\alpha'), \alpha, t)\, d\alpha' dv' \\ &+\frac{1}{N(t)} \int_{A_-} v' \rho(0, v', \alpha', t) \rho(l, v, \alpha, t)\, d\alpha' dv'. \end{aligned} \tag{28}$$

This evolution equation has been simulated and produced reasonable results when compared to the microscopic dynamics, as shown later in Section 5. The major obstacle in using this approach lies in the increased computational complexity associated with evaluating double integrals in (28).

4.3. Model B. Here we assume that the state of a grain boundary is given by four parameters $(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)})$. Although the phase space is larger in this case, the collision rules are simpler than those for the model A. Indeed, the velocity of both the boundary and the junctions with its neighbors to the right and to the left can be uniquely determined via (11) and (4), respectively.

Set $\mathcal{B} := \mathbf{R}_+ \times \mathbf{R}^3$. Suppose that a grain boundary $\big(0, \beta_{(-1)}, \beta, \beta_{(+1)}\big) \in \mathcal{B}$ disappears to the left of the grain boundary $\big(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)}\big) \in \mathcal{B}$ (Figure 6). Clearly, $\beta = \alpha_{(-1)}$ and $\beta_{(+1)} = \alpha$. Further, the collision leads to the formation of the new boundary $\big(l, \beta_{(-1)}, \alpha, \alpha_{(+1)}\big) \in \mathcal{B}$.

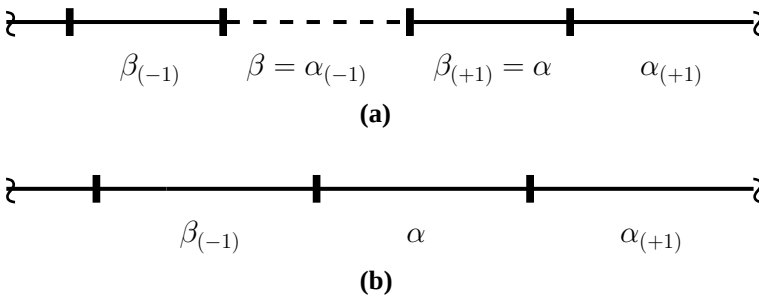


FIGURE 6. Orientations of the grain boundaries (a) before and (b) after a collision. The disappearing boundary is represented by the dashed line.

As before, let $N(t)$ be the number of grain boundaries in the system at time t, denote $N_0 := N(0)$, and suppose that $\rho\left(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)}, t\right)$ represents the number density of states of the grain boundary system at a time t. Then ρ satisfies

$$\int_{\mathcal{B}} \rho\left(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)}, t\right)\, dl\, d\alpha\, d\alpha_{(-1)}\, d\alpha_{(+1)} = N(t). \tag{29}$$

Further, we assume that $\rho(l, x, y, z, t) = \rho(l, z, y, x, t)$ for every $(l, z, y, x) \in \mathcal{B}$ and $t \geq 0$.

The rate at which the grain boundaries are added to an element of the phase space $[l, l+dl] \times \left[\alpha_{(-1)}, \alpha_{(-1)} + d\alpha_{(-1)}\right] \times [\alpha, \alpha + d\alpha] \times \left[\alpha, \alpha_{(+1)} + d\alpha_{(+1)}\right]$ is given by

$$\begin{aligned} W_+ &:= \tfrac{1}{N(t)} \textstyle\int_{\mathcal{B}} \left(2f(s) - f(\alpha) - f(\alpha_{(-1)})\right) \rho\left(0, \alpha_{(-1)}, s, \alpha\right) \rho\left(l, s, \alpha, \alpha_{(+1)}\right)\, ds \\ &+ \tfrac{1}{N(t)} \textstyle\int_{\mathcal{B}} \left(2f(s) - f(\alpha) - f(\alpha_{(+1)})\right) \rho\left(0, \alpha, s, \alpha_{(+1)}\right) \rho\left(l, \alpha_{(-1)}, \alpha, s\right)\, ds. \end{aligned} \tag{30}$$

Similarly

$$\begin{aligned} W_- &:= \tfrac{1}{N(t)} \textstyle\int_{\mathcal{B}} \left(f(\alpha) + f(s) - 2f(\alpha_{(-1)})\right) \rho\left(0, s, \alpha_{(-1)}, \alpha\right) \rho\left(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)}\right)\, ds \\ &+ \tfrac{1}{N(t)} \textstyle\int_{\mathcal{B}} \left(f(\alpha) + f(s) - 2f(\alpha_{(+1)})\right) \rho\left(0, \alpha, \alpha_{(+1)}, s\right) \rho\left(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)}\right)\, ds. \end{aligned} \tag{31}$$

The kinetic equation has the following form

$$\begin{aligned} &\frac{\partial \rho\left(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)}\right)}{\partial t} \\ &+ \left(f\left(\alpha_{(-1)}\right) + f\left(\alpha_{(+1)}\right) - 2f(\alpha)\right) \frac{\partial \rho\left(l, \alpha_{(-1)}, \alpha, \alpha_{(+1)}\right)}{\partial l} = W, \end{aligned} \tag{32}$$

where the collision integral $W = W_+ - W_-$ and W_+ and W_- are given by (30)-(31).

5. Numerical results

Here we find the numerical solutions of the equations (28) and (32) and compare the results with those obtained by simulating the deterministic one-dimensional system of grain boundaries.

We begin by describing the numerical procedure for the model A—the procedure is the same for model B (with minor modifications). First, we construct the initial condition for the number density of states. We fix the initial number of grain boundaries in the system to be $n = 10000$ and supply each model with a random

input data in the form of n orientation parameters α and n randomly distributed lengths of the boundaries l. Velocities are then computed by the rules given in (4).

Note that the region occupied by the system in the phase space continuously grows with time in the l-direction. In order to keep the numerical procedure simple and the size of the simulation small, we introduce an artificial constraint on a grain boundary length by assuming that l cannot exceed some (relatively large) $L > 0$. The drawback of imposing this constraint is that the accuracy of the simulation will be affected for large times when the system can grow beyond the computational domain.

Next we discretize the domain $\Omega := [0, L] \times [-1, 1] \times [0, 2]$ in the phase space by using a uniform mesh with $n_\alpha = n_v = 20$ and $n_l = 100$ discretization points in the α-, v-, and l-directions, respectively. Each cell of this discretization also serves as a "bin" containing some states of the randomly generated data; by counting the number of states in each bin we obtain the initial condition on the number density of states.

We discretize equation (28) by using an explicit upwind scheme in which the spatial derivative is discretized as $v\frac{\partial \rho}{\partial l}(l, v, \alpha, t) \sim \frac{v}{2}(\rho(l+dl, v, \alpha, t) - \rho(l-dl, v, \alpha, t)) - \frac{|v|}{2}(\rho(l + dl, v, \alpha, t) - 2\rho(l, v, \alpha, t) + \rho(l - dl, v, \alpha, t))/dl$ for all interior points of Ω; we use a forward difference scheme for the boundary $l = 0$ and a backward difference scheme for the boundary $l = L$. The collision integral is then computed by calculating lower-right Riemann sums for the admissible pairs of (v, α) as specified by the sets A_- and A_+.

In Figures 7-9 we present the results of the numerical experiments. For three different choices of the energy functional, we compare the statistics obtained via simulations of the deterministic system with those obtained by numerically solving the equation (28) of model A.

Although a very good agreement exists for the distributions of both lengths and orientations, the deviation between the corresponding distributions of the velocities becomes significant after some time. There are several factors that can contribute to such behavior. Some factors may be numerical in nature (e.g. there are discretization errors), some may be due to the modeling assumptions that are too restrictive (the development of correlations that are not accounted for in the model), and some may be inherent to the inevitable loss of information when passing from the deterministic to the effective kinetic model (Boltzmann equation does not conserve the total length of all grain boundaries in the system, the number of state variables may be too small, etc.).

The results obtained using the model B agree very well with the deterministic simulations for all times for which the simulations were performed (Figure 10). A drawback of the Model B as compared to the model A, is that the equation (32) requires more variables. While this may not be an issue in the one-dimensional example considered here, there may be significant differences between the two models in the computational power needed to describe grain boundaries in the real two- or three-dimensional systems.

6. Discussion

In this work, we have presented a new framework for modeling critical events in microstructure evolution and analyzed its capabilities by applying it to a simplified model, originally introduced in [17]. The model is specifically designed to target the evolution of triple junctions during grain growth disregarding mean curvature effects present in the real systems. In [17], [18], we analyzed the stochastic properties of

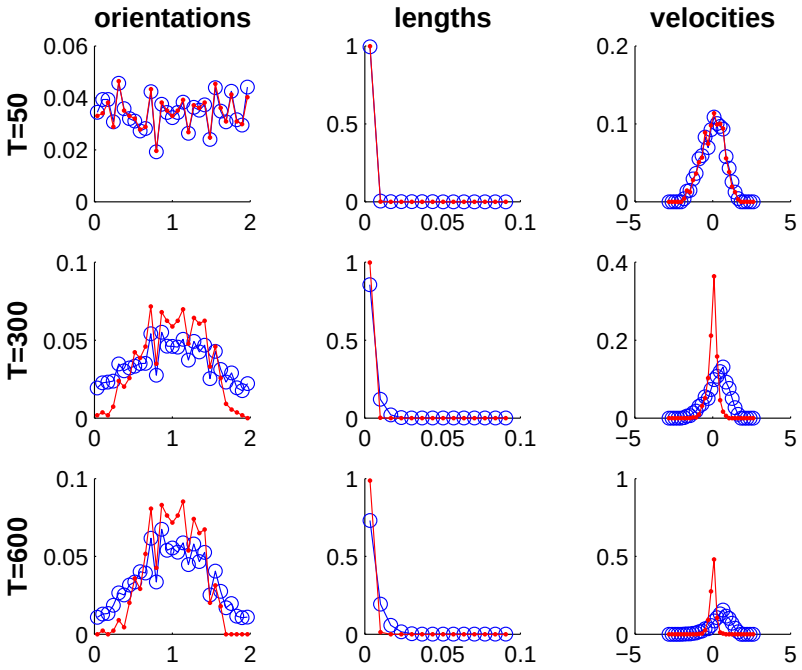


FIGURE 7. Model A. Comparison of the marginal distributions for orientations (left), lengths (center), and velocities (right) when the energy function is given by $f(x) = (x-1)^2$. The deterministic simulations and the solution of the Boltzmann equation (28) are plotted using points and circles, respectively.

this system and discovered that despite its simplicity it exhibits a wide range of complex nonlinear dynamics phenomena, from fractional diffusion to non-identically distributed waiting times. While we have been able to successfully describe some stages of the evolution by means of the random walk theory, the search for a unified and computationally feasible statistical theory is not over. Here we focused on an alternative approach which offers some advantages in describing parts of the system evolution and helps explain some of the stochastic phenomena observed in previous work.

This approach is motivated by the theory of sticky particle dynamics. It has a capability to model critical events more thoroughly through the set of collision rules and hence goes beyond the averaging ideas. The approach proved to be effective in the early stages of the simulation for the model A based on a smaller number of state variables. For larger times, the discrepancy between the kinetic model A and its deterministic counterpart becomes larger. A possible remedy, proposed in this work, is to consider a larger state space kinetic model (model B). The corresponding Boltzmann equation takes into account all local reconfigurations in the grain boundary network and successfully reproduces the distributions during full system lifecycle. This approach, however, may require a significantly larger computations for the higher dimensional problems.

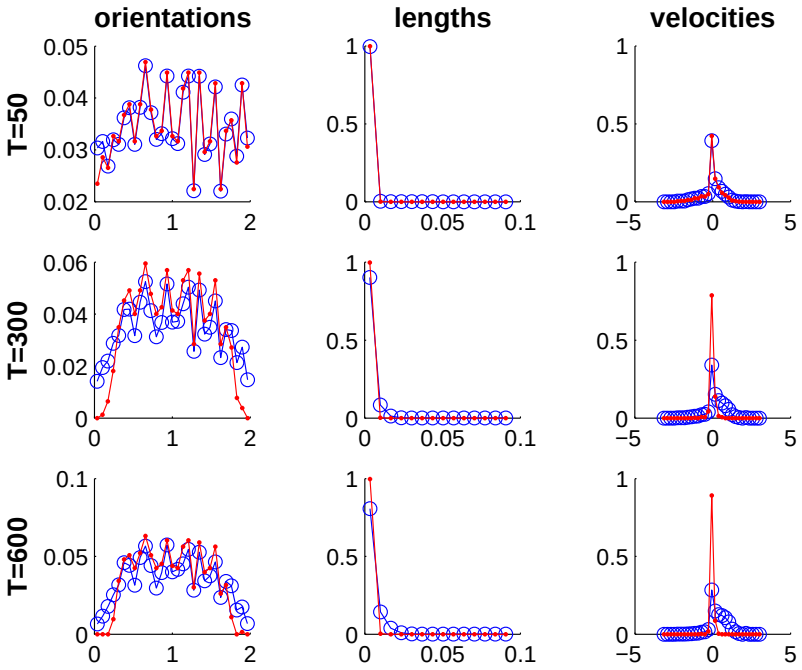


FIGURE 8. Model A. Comparison of the marginal distributions for orientations (left), lengths (center), and velocities (right) when the energy function is given by $f(x) = (x - 0.5)^2(x - 1)^2(x - 1.5)^2$. The deterministic simulations and the solution of the Boltzmann equation (28) are plotted using points and circles, respectively.

7. Acknowledgment

The authors wish to thank their colleagues Eva Eggeling, Gregory Rohrer, and A. D. Rollett.

References

[1] E. M. Lehockey, G. Palumbo, P. Lin, and A. Brennenstuhl. Mitigating intergranular attack and growth in lead-acid battery electrodes for extended cycle and operating life, Metall. Mater. Trans., A Phys. Metall. Mater. Sci., 29 (1998) 7–117.

[2] C. Herring, Surface tension as a motivation for sintering, in The Physics of Powder Metallurgy, W.E. Kingston, ed., MacGraw Hill, New York, (1951), 142-179

[3] Mullins, W. W., A One Dimensional Nearest Neighbor Model of Coarsening, Proceedings of the Calculus of Variations and Nonlinear Material Behavior, Carnegie Mellon University, 1990

[4] Mullins, W.W., Two-dimensional Motion of Idealized Grain Boundaries, J. Appl. Phys., 27 (1956) 900–904.

[5] Mullins, W.W., *Solid Surface Morphologies Governed by Capillarity*, In Metal Surfaces: Structure, Energetics, and Kinetics, ASM, Cleveland, 1963

[6] S. Agmon, A. Douglis and L. Nirenberg, Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions, II, Comm. Pure Appl. Math. 17 (1964) 35–92.

[7] L. Bronsard and F. Reitich, On three-phase boundary motion and the singular limit of a vector-valued Ginzburg-Landau equation, Arch. Rat. Mech. Anal. 124 (1993) 355–379.

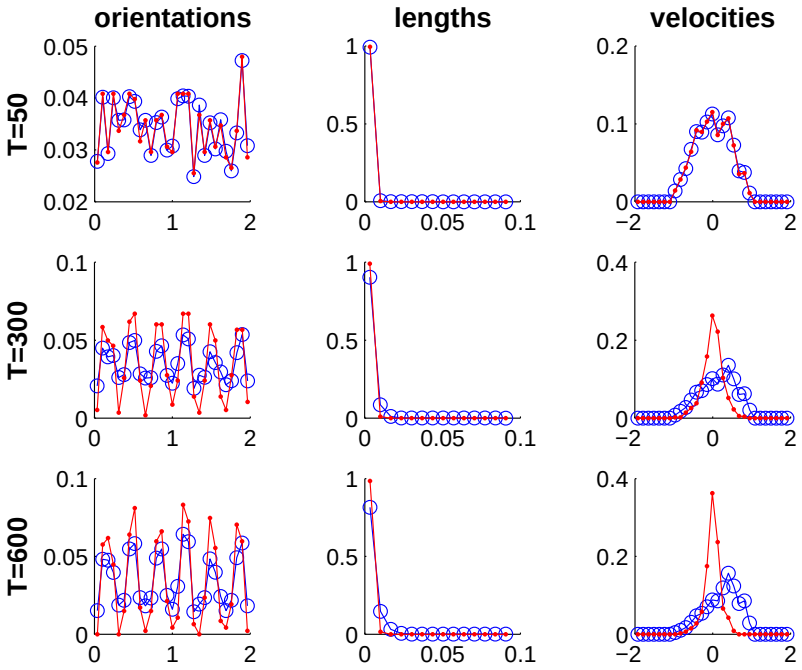


FIGURE 9. Model A. Comparison of the marginal distributions for orientations (left), lengths (center), and velocities (right) when the energy function is given by $f(x) = (\cos(6x\pi) + 1)/4$. The deterministic simulations and the solution of the Boltzmann (28) equation are plotted using points and circles, respectively.

[8] D. Kinderlehrer, J. Lee, I. Livshits, and S. Ta'asan. *Mesoscale simulation of grain growth.* In Continuum Scale Simulation of Engineering Materials(Raabe, D. et al., eds), pages 361–372, Wiley-VCH Verlag, Weinheim, 2004.

[9] D. Kinderlehrer, I. Livshits, G. S. Rohrer, S. Ta'asan, and P. Yu, Mesoscale evolution of the grain boundary character distribution. Recrystallization and Grain Growth, Materials Science Forum 467-470 (2004) 1063–1068.

[10] D. Kinderlehrer, I. Livshits, F. Manolache, G. S. Rohrer, S. Ta'asan, *An approach to the mesoscale simulation of grain growth* In Influences of interface and dislocation behavior on microstructure evolution (Aindow, M. et al., eds), Mat. Res. Soc. Symp. Proc. 652, Y1.5., 2001.

[11] D. Kinderlehrer, I. Livshits, and S. Ta'asan, A variational approach to modeling and simulation of grain growth, SIAM J. Sci. Comput., 28 (2006) 1694–1715.

[12] D. Kinderlehrer and C. Liu, Evolution of grain boundaries, Math. Models and Meth. Appl. Math., 11.4 (2001) 713–729.

[13] R. D. MacPherson and D. J. Srolovitz, The von Neumann relation generalized to coarsening of three-dimensional microstructures, Nature, 446 (2007) 1053–1055.

[14] J. Gruber, thesis, CMU 2007, cf. also G.S. Rohrer, Influence of Interface Anisotropy on Grain Growth and Coarsening," Annual Review of Materials Research 35 (2005) 99-126

[15] V.E. Fradkov, D. Udler, Two-dimensional normal grain growth: topological aspects, Advances in Physics, 43 (1994) 739–789.

[16] H.V.Atkinson, Theories on normal grain growth in pure single phase systems, Acta Metall., 36 (1988) 469–491.

[17] M. Emelianenko, D. Golovaty, D. Kinderlehrer, and S. Taasan, Toward a statistical theory of texture, submitted to SIAM J. Sci. Comput., 2007.

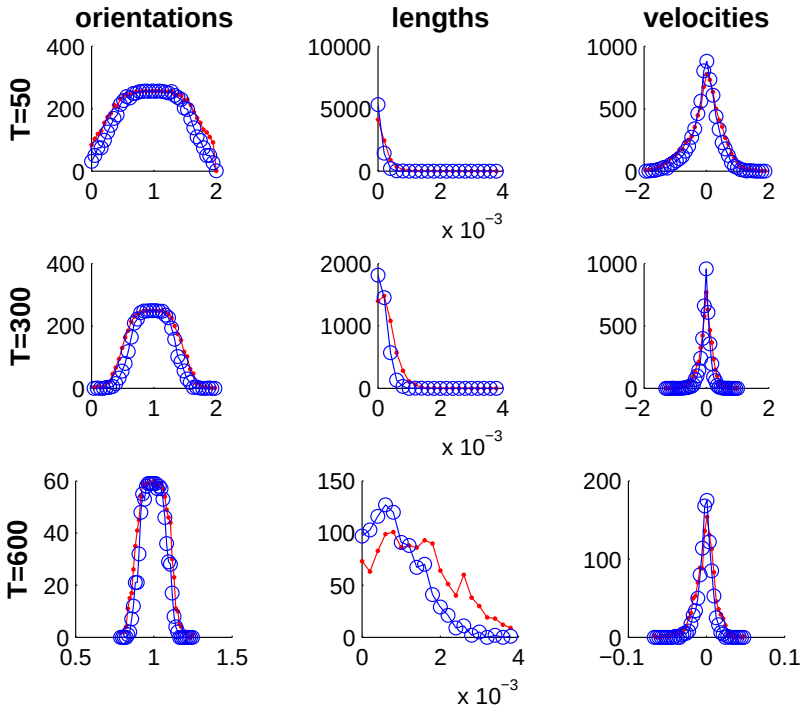


FIGURE 10. Model B. Comparison of the marginal distributions for orientations (left), lengths (center), and velocities (right) when the energy function is given by $f(x) = (x-0.5)^2$. The deterministic simulations and the solution of the Boltzmann equation (32) are plotted using points and circles, respectively.

[18] K. Barmak, M. Emelianenko, D. Golovaty, D. Kinderlehrer, and S. Taasan., *On a statistical theory of critical events in microstructural evolution*, in proceedings of 11th International Symposium on Continuum Models and Discrete Systems, Paris, 30 July - 3 August, 2007 (CMDS11), 2007

[19] Anna de Masi, Enrico Presutti, Mathematical Methods for Hydrodynamic Limits, Springer-Verlag, 1991

[20] Ludwig Boltzmann, Lectures on gas theory, Dover, 1995

[21] D. Benedetto, E. Caglioti, M. Pulvirenti, A kinetic equation for granular media, Math. Mod. and Num. An., 31 (1997) 615–641.

Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213

E-mail: katayun@andrew.cmu.edu

Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030

E-mail: memelian@gmu.edu

Department of Theoretical and Applied Mathematics, The University of Akron, Akron, OH 44325

E-mail: dmitry@math.uakron.edu

Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213

E-mail: davidk@cmu.edu and shlomo@andrew.cmu.edu

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 109–125

A NOTE ON THE CONSTRUCTION OF FUNCTION SPACES FOR DISTRIBUTED-MICROSTRUCTURE MODELS WITH SPATIALLY VARYING CELL GEOMETRY

SEBASTIAN MEIER AND MICHAEL BÖHM

Abstract. We construct Lebesgue and Sobolev spaces of functions defined on a continuous distribution of domains $\{Y_x \subset \mathbb{R}^m : x \in \Omega\}$. The resulting spaces can be viewed as a generalisation of the Bochner spaces $L_p(\Omega; W_q^l(Y))$ for the case that Y depends on $x \in \Omega$. Furthermore, we introduce a Lebesgue space of functions defined on the boundaries $\{\partial Y_x : x \in \Omega\}$. The latter construction relies on a uniform Lipschitz parametrisation of the above collection of boundaries, interpreted as a higher-dimensional manifold. The results are applied to prove existence, uniqueness and upper and lower bounds for a distributed-microstructure model of reactive transport in a heterogeneous porous medium.

Key Words. Lebesgue spaces, Sobolev spaces, distributed-microstructure model, direct integral, reaction–diffusion, homogenisation.

1. Introduction

Transport in porous media is governed by at least two highly different spatial scales: the *pore scale* and the *macroscopic scale*, the latter of which is usually of interest in applications. In cases where two or more transport processes happen simultaneously on highly different time scales, it has been shown by periodic homogenisation that *distributed-microstructure models* (or *two-scale models*) are appropriate [3, 2]. Such models consist of averaged equations describing the fast transport processes and of local microscopic cell problems accounting for the slow transport. The most studied example is flow in fissured media [1, 25].

¿From a mathematical point of view, these models are interesting due to the non-standard coupling of the equations and the unusual choice of solution spaces. In [25], the authors show that the variational formulation of a distributed-microstructure model with a cell geometry that varies at different points of the medium naturally leads to function spaces of the form $L_2(\Omega; H^1(Y_x))$ where Y_x is another domain depending on $x \in \Omega$. The construction of such spaces and particularly of their trace spaces is quite intricate and it is the major aim of this paper.

We briefly recall the model from [25] and how a variational formulation is derived. If $\Omega \subset \mathbb{R}^n$ is the macroscopic flow region, then at each $x \in \Omega$ the local geometry is described by a solid matrix block $Y_x \subset Y \subset \mathbb{R}^n$ surrounded by the pore $Y \setminus \overline{Y}_x$. The domain Y_x can depend on the macroscopic space coordinate $x \in \Omega$ in order to account for a heterogeneous medium. For $x \in \Omega$, $y \in Y_x$ and $t \geq 0$, let $u(x,t)$ be the fluid density in the pore space and $U(x,y,t)$ that in the matrix blocks. The

2000 *Mathematics Subject Classification.* 46E30, 46E35, 35K57, 35B27.

model equations consist of the (averaged) mass balance of fluid within the pores[1]

(1a)
$$\frac{\partial}{\partial t}(a(x)u) - \mathrm{div}_x(A(x)\nabla_x u) = \frac{1}{|Y|}\int_{\partial Y_x} k(\gamma_x U(x,y,t) - u(x,t))\cdot\nu \, d\sigma_y, \quad x \in \Omega,\ t > 0,$$

where $\gamma_x U(t,x,y)$ denotes the trace of U at $y \in \partial Y_x$, and a family of local mass balances in the matrix blocks parameterised by $x \in \Omega$,

$$\frac{\partial}{\partial t}(b(x)U) - \mathrm{div}_y(B(x)\nabla_y U) = 0, \qquad x \in \Omega,\ y \in Y_x,\ t > 0. \tag{1b}$$

The exchange condition reads

$$-B(x)\nabla_y U \cdot \nu_x = k(\gamma_x U(x,y,t) - u(x,t)), \qquad x \in \Omega,\ y \in \partial Y_x,\ t > 0. \tag{1c}$$

Following [25], a variational formulation of (1) is given as follows. Let $V := L_2(\Omega; H^1(Y_x))$ be an anisotropic Sobolev space (see Def. 4). We look for a pair of functions $u \in L_2(0,T;H^1(\Omega))$ and $U \in L_2(0,T;V)$ satisfying (1a) in the usual weak sense and

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega\int_{Y_x} bU\Psi \, dy\, dx + \int_\Omega\int_{Y_x} B\nabla_y U \cdot \nabla_y \Psi \, dy\, dx + \int_\Omega\int_{\partial Y_x} k(\gamma_x U - u)\,\gamma_x\Psi \, d\sigma_y \, dx = 0 \qquad \forall \Phi \in V.$$

In [25], the authors prove that the system (1) is wellposed in the above sense. However, a systematic discussion of the properties spaces of the form $L_2(\Omega; H^1(Y_x))$ is missing. Moreover, the cell boundaries Γ_x need to have some regularity with respect to $x \in \Omega$ in order to justify terms of the form

$$\int_\Omega\int_{\partial Y_x} \gamma_x U \, \gamma_x \Psi \, d\sigma_y \, dx.$$

It is the aim of this paper to fill this gap by constructing general spaces $L_p(\Omega; W_q^l(Y_x))$ and $L_p(\Omega; L_q(\partial Y_x))$ and proving some elementary properties of them like separability and reflexivity. While for the former space, it is sufficient that the higher-dimensional set $Q := \cup_{x\in\Omega}(\{x\} \times Y_x)$ is Lebesgue measurable, it turns out that for the latter space of functions defined on a family of cell *boundaries*, the situation is more intricate. We construct a uniform parametrisation of the cell boundaries ∂Y_x under quite general conditions on the geometry. With this framework at hand, objects like the *distributed trace operator* $\gamma U(x,y) := \gamma_x U(x,\cdot)(y)$ are easily constructed. Afterwards, the results are applied to a semilinear two-scale reaction–diffusion system, which has also been discussed in [17] under stronger restrictions on the cell geometry. Modifying techniques from [14, 9], we prove boundedness, existence and uniqueness of weak solutions.

We mention some related work for constant microstructure: The analysis of a similar two-scale reaction–diffusion system has been shown in [10]. Homogenisation results for a general diffusion–convection–reaction–adsorption system can be found in [12, 13]. For numerical approaches to two-scale models, see [21, 1, 18]. A huge list of further references is also given in [11]. We emphasise that in the present paper and in all of the above cited work, a change of the microstructure *w.r.t. time* is not considered. For homogenisation and two-scale models with evolving microstructure, we refer to [22, 16].

This paper is organised as follows. In section 2, we discuss function spaces on cell *domains*. Function spaces on the cell *boundaries* are treated in section 3. In

[1]The model (1) corresponds to the *regularised-microstructure* case in [25].

section 4, we apply the results to prove well-posedness of a two-scale model for reactive transport.

2. Spaces of functions defined in the cell

We construct spaces $L_p(\Omega; L_q(Y_x))$ and $L_p(\Omega; W^l_q(Y_x))$ of functions defined on a family of bounded domains $\{Y_x \subset \mathbb{R}^m : x \in \Omega \subset \mathbb{R}^n\}$. As our basic tool, we use the $(n+m)$-dimensional Lebesgue measure of the domain $Q = \cup_{x\in\Omega}(\{x\} \times Y_x)$. The corresponding spaces of Bochner integrable functions are recovered as special cases if $Y_x \equiv Y$. For general information on the Bochner integral, see [28, 15].

Some of the following results are well-known for the special case that $\Omega = [0, T]$ and x is the time variable. In this case, these spaces are widely used when dealing with free-boundary problems or PDEs on noncylindrical domains. See [19] for similar definitions and further references.

2.1. The Lebesgue space $L_p(\Omega; L_q(Y_x))$. Let $\Omega \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be bounded domains. For each $x \in \Omega$, let $Y_x \subset Y \subset \mathbb{R}^m$ be another domain such that

$$Q := \Omega \times Y_x := \bigcup_{x\in\Omega} (\{x\} \times Y_x) \subset \mathbb{R}^{n+m}$$

is measurable with respect to the $(n+m)$-dimensional Lebesgue measure. If no further restrictions are given, then $p \in [1, \infty]$ and $q \in [1, \infty)$ are given exponents and p' and q' are the dual exponents defined by $1/p + 1/p' = 1$ and $1/q + 1/q' = 1$. The case $q = \infty$ is not considered in this paper.

Definition 1 (The space $L_{p,q}(Q)$)**.**

(1) We define the Banach space

$$\begin{aligned} L_{p,q}(Q) &\equiv L_p(\Omega; L_q(Y_x)) \\ &:= \{u \in L_p(\Omega; L_q(Y)) : \ u(x, \cdot) = 0 \text{ on } Y \setminus Y_x \quad \text{for a.e. } x \in \Omega\} \end{aligned}$$

with the norm

$$\|u\|_{L_{p,q}(Q)} := \begin{cases} (\int_\Omega \|U(x)\|^p_{L_q(Y_x)}\, dx)^{1/p}, & p < \infty, \\ \operatorname{ess\,sup}_{x\in\Omega} \|U(x)\|_{L_q(Y_x)}, & p = \infty. \end{cases}$$

(2) In the case $p = q = 2$, we define the Hilbert space $L_{2,2}(Q)$ with the scalar product

$$(u,\, v)_{L_{2,2}(Q)} := \int_\Omega (u(x),\, v(x))_{L_2(Y_x)}\, dx.$$

Remark. Since $q < \infty$, the function $x \mapsto \|u(x, \cdot)\|_{L_q(Y_x)}$ is measurable by Fubini's theorem. Thus, the space $L_{p,q}(Q)$ is well-defined. As a closed subspace of $L_p(\Omega; L_q(Y))$, it is also complete.

Proposition 2 (Properties of the space $L_{p,q}(Q)$)**.**

(1) (Hlder's inequality) For all $u \in L_p(\Omega; L_q(Y_x))$ and $v \in L_{p'}(\Omega; L_{q'}(Y_x))$, it holds

$$\int_\Omega \int_{Y_x} u(x, y)v(x, y)\, dy\, dx \le \|u\|_{L_{p,q}(Q)} \|v\|_{L_{p',q'}(Q)}.$$

(2) Let $p < \infty$. Then $L_{p,p}(Q)$ is isometrically isomorph to $L_p(Q)$.

(3) Let $p < \infty$. Then the simple functions as well as the continuous functions on $\overline{Q}$ are dense in $L_{p,q}(Q)$. In particular, $L_{p,q}(Q)$ is separable.

Proof. Part (1) is obtained straightforwardly from the standard Hlder's inequalities in the spaces $L_p(\Omega)$ and $L_q(Y_x)$.

Part (2) follows via extension by zero from the corresponding fact for the Bochner space $L_p(\Omega; L_p(Y))$. A proof of the latter result can be found in [8], pp. 196ff.

(3) By definition, $f \in L_{p,q}(Q)$ is Bochner-integrable as a function $f : \Omega \to L_q(Y)$. Therefore we can approximate f by simple functions $f_k : \Omega \to L^q(Y)$ via

$$f_k(x) := \sum_{i=1}^{m_k} \alpha_i^k \mathbb{1}_{E_i^k}(x), \quad \alpha_i^k \in L^q(Y),\ E_i^k \subset \Omega \text{ measurable for } i,k \in \mathbb{N},$$

such that $f_k \to f$ in $L_p(\Omega; L_q(Y))$ for $k \to \infty$. Moreover, each α_i^k can be approximated by

$$\alpha_i^{kl}(y) := \sum_{j=1}^{n_l} \beta_{ij}^{kl} \mathbb{1}_{D_{ij}^{kl}}(y), \quad \beta_{ij}^{kl} \in \mathbb{R}, \quad D_{ij}^{kl} \subset Y \text{ measurable for } j,l \in \mathbb{N}.$$

Then one easily verifies that the simple functions on Q given by

$$f^{kl}(x,y) := \sum_{i=1}^{m_k} \sum_{j=1}^{n_l} \beta_{ij}^{kl} \mathbb{1}_{E_i^k \times (D_{ij}^{kl} \cap Y_x)}(x,y)$$

approximate f in $L^{p,q}(Q)$ for $k,l \to \infty$. In order to prove the density of $C(\overline{Q})$ in $L_{p,q}(Q)$, it suffices to approximate simple functions $f : Q \to \mathbb{R}$. Thus, we can assume that $f \in L^r(Q)$ for every $r \geq 1$. Since $C(\overline{Q})$ is dense in $L_{\max\{p,q\}}(Q)$, the result follows. □

Proposition 3 (Characterisation of the dual space)**.** Let $p \in [1,\infty)$ and $q \in (1,\infty)$. The operator

$$\langle J(f),\, g\rangle := \int_\Omega \int_{Y_x} f(x,y) g(x,y)\, dy\, dx, \qquad g \in L_{p,q}(Q),\ f \in L_{p',q'}(Q),$$

is an isometric isomorphism $J : L_{p',q'}(Q) \to [L_{p,q}(Q)]'$. In particular, $L_{p,q}(Q)$ is reflexive for $p > 1$.

Remark. The case $q = 1$ is not covered since the space $L_{p',\infty}(Q)$ has not been defined.

Proof. Let $f \in L_{p',q'}(Q)$. By Hlder's inequality, $J(f)$ is well-defined and $\|J(f)\| \leq \|f\|_{p',q'}$. Moreover, by the fundamental lemma of calculus of variations, J is injective.

Let $F \in [L_{p,q}(Q)]'$ be given. We have to show that an $f \in L_{p',q'}(Q)$ exists with

$$F = J(f) \qquad \text{and} \qquad \|f\|_{p',q'} \leq \|F\|.$$

We reduce the statement to the cylindrical situation in the following way: Define

$$\tilde{J} : L_{p'}(\Omega; L_{q'}(Y)) \to [L_p(\Omega; L_q(Y))]', \qquad \langle \tilde{J}(f),\, g\rangle := \int_\Omega \int_Y f g\, dy\, dx.$$

In this case, it is known that $\tilde{J}$ is an isometric isomorphism. The proof, which is very similar to the real-valued case $L_p(\Omega)$, can be found in [4]. Define $\lambda_{p,q} : L_p(\Omega; L_q(Y)) \to L_{p,q}(Q)$ to be the linear restriction operator and let $\lambda'_{p,q}$ be its dual. Both operators have norm 1, and the right-inverse of $\lambda_{p,q}$ is the extension by zero, which we denote by $\gamma_{p,q} : L_{p,q}(Q) \to L_p(\Omega; L_q(Y))$. Hence, $\gamma_{p,q} \circ \lambda_{p,q} = \mathrm{id}$. Let

$$f := \lambda_{p',q'} \circ \tilde{J}^{-1} \circ \lambda'_{p,q} \circ F \in L_{p',q'}(Q).$$

Then $\|f\|_{p',q'} \le \|F\|$ and, for any $g \in L_{p,q}(Q)$,

$$\begin{aligned}\langle J(f),\, g\rangle &= \left\langle J \circ \lambda_{p',q'} \circ \tilde{J}^{-1} \circ \lambda'_{p,q} \circ F,\, g\right\rangle \\ &= \int_\Omega \int_{Y_x} (\lambda_{p',q'} \circ \tilde{J}^{-1} \circ \lambda'_{p,q} \circ F)\, g\, dy\, dx \\ &= \int_\Omega \int_{Y} (\tilde{J}^{-1} \circ \lambda'_{p,q} \circ F)\, \gamma_{p,q} g\, dy\, dx \\ &= \left\langle \lambda'_{p,q} \circ F,\, \gamma_{p,q} \circ g\right\rangle \\ &= \left\langle F,\, (\lambda_{p,q} \circ \gamma_{p,q}) g\right\rangle \\ &= \langle F,\, g\rangle.\end{aligned}$$

Hence $J(f) = F$, which proves the Proposition. □

2.2. The Sobolev space $L_p(\Omega; W^l_q(Y_x))$. In the following, let $l \in \mathbb{N}$. For a multi-index $\alpha \in \{0, \dots, l\}^m$, let $\partial_\alpha u(x, y)$ denote the α-th derivative w.r.t. y.

Definition 4. We define the Banach space

$$W^{0,l}_{p,q}(Q) \equiv L_p(\Omega; W^l_q(Y_x)) := \{u \in L_{p,q}(Q) :\ \partial^\alpha u \in L_{p,q}(Q)\ \forall |\alpha| \le l\}$$

with the norm

$$\|u\|_{W^{0,l}_{p,q}(Q)} := \sum_{|\alpha|\le l} \|\partial_\alpha u\|_{L_{p,q}(Q)}.$$

In the case $p = q = 2$, $W^{0,l}_{2,2}(Q)$ is a Hilbert space with the scalar product

$$(u,\, v)_{W^{0,l}_{2,2}(Q)} := \sum_{|\alpha|\le l} (\partial_\alpha u,\, \partial_\alpha v)_{L_{2,2}(Q)}\,.$$

Remark. The proof that $W^{0,l}_{p,q}(Q)$ is complete reduced to the completeness of $L_{p,q}(Q)$ by the analogous argument as for standard Sobolev spaces.

Proposition 5 (Properties of $W^{0,l}_{p,q}(Q)$)**.** The space $W^{0,l}_{p,q}(Q)$ is separable if $p < \infty$ and reflexive if $p, q \in (1, \infty)$.

Proof. In order to prove separability, define the linear bounded operator

$$T : X := W^{0,l}_{p,q}(Q) \to \prod_{|\alpha|\le l} L_{p,q}(Q), \qquad f \mapsto (\partial_\alpha f)_{|\alpha|\le l}.$$

Then $T(f)$ can be estimated from above and below by the norm $\|f\|_X$. Thus, X is isomorph to $T(X)$. Since $C(\overline{Q})$ is dense in $L_{p,q}(Q)$ for $p < \infty$, the subspace $T(X)$ is separable and, hence, also X.

Finally, for $p, q \in (1, \infty)$, $L_{p,q}(Q)$ is reflexive by Prop. 3. Since X is a Banach space, $T(X)$ is a closed subspace of $L_{p,q}(Q)$ and is therefore also reflexive. □

Remark. As already mentioned in [25], for the special case $q = 2$ and $p < \infty$, the space $W^{0,l}_{p,2}(Q)$ can alternatively be constructed as a *direct integral of Hilbert spaces*. See [7, 26] for abstract definitions, or [6] for the special case that $n = 1$ and $\Omega = (a, b)$. For details on how the construction works for the space $W^{0,l}_{p,2}(Q)$, the reader is referred to [16].

3. Spaces of functions defined on the cell boundary

In order to give a meaning to integrals of the form $\int_\Omega \int_{\partial Y_x} u(x,y)\, d\sigma_y\, dx$, we need to define spaces of functions defined on the cell boundary ∂Y_x. Here the situation is more complicated since we need to parameterise the whole collection of cell boundaries $\{\Gamma_x : x \in \Omega\}$. For a concise presentation, we exclude some (degenerate) geometries that could perhaps be treated with more technical effort.

3.1. Parametrisation of the cell boundaries. We assume that the Lebesgue measure of the cells $Y_x \subset \mathbb{R}^m$ is uniformly bounded from below, i.e., there exists a constant $c > 0$ such that $|Y_x| \geq c$ for all $x \in \Omega$. Let us introduce a parametrisation of

$$\Sigma := \overline{\Omega} \times \partial Y_x := \bigcup_{x \in \overline{\Omega}} (\{x\} \times \partial Y_x) \subset \partial Q.$$

Note that ∂Q is not smoother than Lipschitz, even if Y_x and Ω have smooth boundaries. In general, the normal will be multi-valued at points $(x,y) \in \partial\Omega \times \partial Y_x$. In order to avoid technicalities, we therefore assume that Q can be extended to a bounded Lipschitz domain $Q_1 \supset Q$ (cf. figure 2 on the left) and construct a parameterisation of $\Sigma_1 := \partial Q_1 \supset \Sigma$.

We recall the definition of a Lipschitz boundary (cf. [27], e.g.). At any given point $(x_0, y_0) \in \Sigma_1$, there exists a neighbourhood U of (x_0, y_0) such that, after an Euclidean coordinate transform, $\Sigma_1 \cap U$ is the graph of a (locally) Lipschitz continuous function $g : \mathbb{R}^{n+m-1} \to \mathbb{R}$ and $Q_1 \cap U$ lies on one side of the graph. See figure 1. More precisely, there exists an orthonormal basis $\{v_j\}_{j=1,\dots,n+m}$, a number $r > 0$, such that with the notation $\xi' = (\xi_1, \dots, \xi_{n+m-1})$, the bijective coordinate transform

$$\Psi^{-1} : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}, \quad \xi \mapsto (x,y) = (x_0,y_0) + \sum_{i=1}^{n+m-1} \xi_j v_j + (\xi_{n+m} + g(\xi'))v_{n+m},$$

the cube $W^{n+m} = (-r,r)^{n+m}$, $r > 0$, and the open set

$$U = \Psi^{-1}(W^{n+m}) \subset \mathbb{R}^{n+m},$$

it holds for all $(x,y) \in U$

$$(x,y) \in U \cap \Sigma_1 \iff \xi_{n+m} = g(\xi'), \tag{2}$$

$$(x,y) \in U \cap Q_1 \iff 0 < \xi_{n+m} - g(\xi') < r, \tag{3}$$

$$(x,y) \in U \setminus \overline{Q_1} \iff -r < \xi_{n+m} - g(\xi') < 0. \tag{4}$$

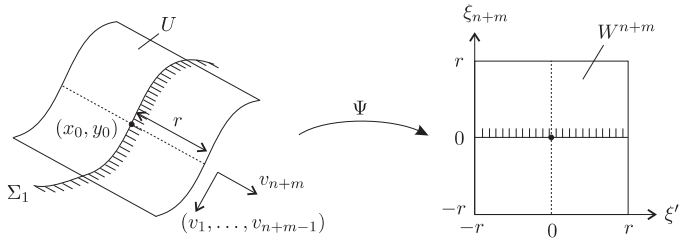


FIGURE 1. Parametrisation of the $(n+m-1)$-dimensional manifold Σ_1.

The set of unit normal vectors directed outward is defined (independently of the parametrisation) as

$$\nu_\Sigma(x,y) := (1+|\nabla_{\xi'} g(\xi')|^2)^{-1/2} \left(\sum_{j=1}^{n+m-1} \partial_j g(\xi') v_j - v_{n+m} \right), \qquad (x,y) \in \Sigma_1. \tag{5}$$

Here, $\partial_j g$ and $\nabla_{\xi'} g$ can be multi-valued and refer to the Clarke gradient (see [5]). In general, at some given point $(x,y) \in \Sigma_1$ the normal $\nu_\Sigma(x,y)$ is a convex set of vectors, but for almost every (x,y) it is a singleton.

We recall how the integral over the $(n+m-1)$-dimensional manifold Σ is defined. Since Σ is compact, there exist finitely many of such parameterisations (x_0^k, y_0^k), U^k, g^k, $\{v_j^k\}_{j=1,\dots,n+m}$ such that $\Sigma \subset \cup_{k=1}^N U^k$. Let $(\eta^k)_{k=1,\dots,N}$ be a partition of unity subordinated to the U^k. We define the integral of a function $f : \Sigma \to \mathbb{R}$ to be

$$\int_\Sigma f d\sigma_{x,y} := \sum_{k=1}^N \int_{W^{n+m-1}} (\eta^k f) \left((\Psi^k)^{-1}(\xi', 0) \right) \sqrt{1 + |\nabla_{\xi'} g^k(\xi')|^2}\, d\xi'.$$

Note that the integral is single-valued, as the set of points where the g^k are not differentiable is of measure zero (Rademacher's theorem).

Next we construct a parametrisation of the cell boundaries $\{\partial Y_x : x \in \Omega\}$ from the above one. Cf. figure 2 in the center, we define the normal ν_x at the cell boundary ∂Y_x to be the (normalised) projection P_y of ν_Σ onto the subspace $\{x = 0\}$ spanned by the vectors $e_{n+1}, \dots, e_{n+m}$, with other words,

$$\nu_x(y) := \frac{\mathrm{P}_y \nu_\Sigma(x,y)}{|\mathrm{P}_y \nu_\Sigma(x,y)|}, \qquad (x,y) \in \Sigma. \tag{6}$$

For this definition to make sense, it is necessary that the projection is nonzero. It can be shown that this condition is also *sufficient* for constructing a parametrisation of the cell boundaries. This is the purpose of the following Lemma.

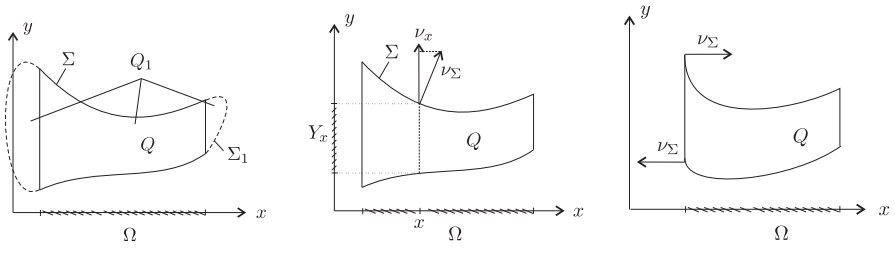


FIGURE 2. Left: The extension Σ_1 of the boundary. Center: Unit normal ν_Σ at Σ and cell normal ν_x. Right: Example of a geometry that does *not* satisfy condition (7).

Lemma 6. Assume that

$$\mathrm{P}_y \nu_\Sigma(x,y) \not\equiv 0 \qquad \forall (x,y) \in \Sigma. \tag{7}$$

Then the cells Y_x are bounded Lipschitz domains with unit normals at ∂Y_x given by (6) and the parametrisation in each point $(x_0, y_0) \in \Sigma$ can be chosen such that

$$v_j = e_j, \qquad j = 1, \dots, n.$$

Moreover, there exist positive constants c and C such that

$$c \le |Y_x|, |\partial Y_x| \le C \qquad \text{for all } x \in \Omega. \tag{8}$$

Remark. It is important that (7) is valid for all (x, y) in the *closed* set $\Sigma = \overline{\Omega} \times \partial Y_x$. By this assumption, we exclude some degenerate cases; see figure 2 on the right.

Proof. Let $(x_0, y_0) \in \Sigma$ be an arbitrary point and assume that a local parametrisation in a neighbourhood U is given as above. We construct a new parametrisation as required. We write (2) in (x, y)-coordinates:

$$v_{n+m} \cdot (x,y) = g(v_1 \cdot (x,y), \dots, v_{n+m-1} \cdot (x,y)) \qquad \forall (x,y) \in U \cap \Sigma_1,$$

or

$$G(x,y) := g(v_1 \cdot (x,y), \dots, v_{n+m-1} \cdot (x,y)) - v_{n+m} \cdot (x,y) = 0 \qquad \forall (x,y) \in U \cap \Sigma_1.$$

We have, according to (5) and (7),

$$\begin{aligned} \nabla_y G(x,y) &= \sum_{j=1}^{n+m-1} \partial_j g(\xi') \mathrm{P}_y v_j - \mathrm{P}_y v_{n+m} \\ &= (1 + |\nabla_{\xi'} g(\xi')|)^2)^{1/2} \; P_y \nu_\Sigma(x,y) \;\not\equiv 0 \qquad \forall (x,y) \in U \cap \Sigma_1. \end{aligned}$$

We can therefore apply the implicit function theorem (see [5] for a Lipschitz version) and resolve the equation near (x_0, y_0) with respect to a vector lying in the subspace $\{x = 0\}$ spanned by $e_{n+1}, \dots, e_{n+m}$. More precisely, there exists an ONB $w_1, \dots, w_n$ of $\mathbb{R}^n$ and a (possibly smaller) cube $\tilde{W}^{n+m} = (-\tilde{r}, \tilde{r})^{n+m}$ and a Lipschitz continuous function $\tilde{g} : \mathbb{R}^{n+m-1} \to \mathbb{R}$ such that, with the coordinate transform

$$\Psi^{-1} : \tilde{W}^{n+m} \to \mathbb{R}^{n+m}, \qquad (x,\zeta) \mapsto (x_0, y_0) + \left(x, \sum_{i=1}^{m-1} \zeta_i w_i + (\zeta_m + \tilde{g}(x, \zeta')) w_m \right),$$

(where $\zeta' := (\zeta_1, \dots, \zeta_{n-1})$), it holds for all $(x, y) \in \Psi^{-1}(\tilde{W}^{n+m}) = \tilde{U}$,

$$\zeta_m = \tilde{g}(x, \zeta'), \qquad \Longleftrightarrow \qquad (x,y) \in \Sigma_1 \cap \tilde{U},$$

with other words, since $\tilde{U} = (x_0 - r, x_0 + r) \times \tilde{U}_x$, where $\tilde{U}_x$ is a neighbourhood of $y_0 \in \partial Y_x$,

$$\zeta_m = \tilde{g}(x, \zeta'), \qquad \Longleftrightarrow \qquad y \in \partial Y_x \cap \tilde{U}_x.$$

It follows that, with $\tilde{g}_x(y) := \tilde{g}(x, y)$, $(x, y) \in \Omega \times W^m$, and the new transformation

$$\tilde{\Psi}_x^{-1} = \tilde{\Psi}^{-1}(x, \cdot) : W^m \to \tilde{U}_x, \qquad \zeta \mapsto y_0 + \sum_{i=1}^{m-1} \zeta_i w_i + (\zeta_m + \tilde{g}_x(\zeta')) w_m$$

we have found a parametrisation of ∂Y_x. If the neighbourhood $\tilde{U}$ has been chosen small enough, then the corresponding conditions (2)–(4) are satisfied.

The estimates (8) are now obvious since, due to the compactness of Σ, the whole collection of cells can be covered by a finite number of maps Ψ^k, $k = 1, \dots, N$. □

Motivated by the preceding Lemma, we define:

Definition 7 (Regular family of cells)**.** The family $\{(Y_x, \Gamma_x) : \ x \in \Omega\}$ is called a regular family of cells if $\Omega \subset \mathbb{R}^n$ is a bounded Lipschitz domain and it holds:

(1) For each $x \in \Omega$, $Y_x \subset \mathbb{R}^m$ is a bounded domain such that $Q := \Omega \times Y_x \subset \mathbb{R}^{n+m}$ is measurable. There exists a constant $c > 0$ such that $|Y_x| \geq c$ for all $x \in \Omega$.
(2) For each $x \in \Omega$, Γ_x is a measurable subset of ∂Y_x.
(3) $\Sigma := \partial Q \subset \mathbb{R}^{n+m}$ can be extended to a boundary of a Lipschitz domain $Q_1 \supset Q$ in $\mathbb{R}^{n+m}$, such that the convex set of unit normals ν_Σ at ∂Q_1 satisfies

$$\mathrm{P}_y \nu_\Sigma(x, y) \not\ni 0 \qquad \forall (x, y) \in \Sigma.$$

Let $\{(Y_x, \Gamma_x) : \ x \in \Omega\}$ be a regular family of cells. Remember that we want to define the integral of a function over Γ_x such that it is measurable w.r.t. x. W.l.o.g., we assume that the parametrisation $\{x_0^k, y_0^k, U^k, g^k\}_{k=1,\dots,N}$ has already the form as constructed in the proof of Lemma 6. We define the corresponding transformations Ψ^k such that $\Sigma \subset \cup_{k=1}^N U^k$. Clearly, the x-intersections $U_x^k \subset \mathbb{R}^m$ cover Γ_x for every $x \in \Omega$ and the $\eta_x^k = \eta^k(x, \cdot)$ can be chosen as a partition of unity subordinated to the U_x^k. Thus, we can define:

Definition 8 (Integral over Γ_x)**.** Let $\{(Y_x, \Gamma_x) : \ x \in \Omega\}$ be a regular family of cells. For a function $f : \Sigma \to \mathbb{R}$ which is integrable with respect to the $(n+m-1)$-dimensional Lebesgue measure, we define, for almost every $x \in \Omega$,

$$\int_{\Gamma_x} f(x, y) d\sigma_y := \sum_{k=1}^N \int_{W^{m-1}} (\eta^k f)\left(x, (\Psi_x^k)^{-1}(\zeta')\right) \sqrt{1 + |\nabla_{\zeta'} g_x^k(\zeta')|^2}\, d\zeta'.$$

Remark.

(1) By our assumption on the geometry, jumps in the cell structure along the x-coordinate are excluded. We can account for at least finitely many jumps by a simple modification: Assume that Ω is decomposed into $M \geq 1$ disjoint domains such that $\bar{\Omega} = \cup_{i=1}^M \bar{\Omega}_i$. We now adopt the Lipschitz and geometric conditions for each of the domains Ω_i. The integral is then constructed as the sum of all integrals over Ω_i.
(2) An important special case is given if $\Omega = (a, b)$ is one-dimensional and $t := x \in (a, b)$ represents the time variable. Then the cells $\{Y_t : \ t \in (a, b)\}$, describe a time-dependent domain. In this case, the normal velocity of the interface at $(t, y) \in Q = (a, b) \times \Gamma_t$ is given by

$$w_\Gamma(t, y) = \frac{\mathrm{P}_t \nu_\Sigma(t, y)}{|\mathrm{P}_y \nu_\Sigma(t, y)|}, \qquad \text{for a.e. } t \in (a, b),\ y \in Y_t,$$

where P_t is the projection to $\{y = 0\}$. See also figure 2 in the center. The condition that $\{(Y_t, \Gamma_t) : \ t \in (a, b)\}$ is a regular family of cells guarantees that w_Γ is well-defined almost everywhere.

3.2. The space $L_p(\Omega; L_q(\Gamma_x))$. In what follows, let $\{(Y_x, \Gamma_x) : x \in \Omega\}$ be a regular family of cells.

Proposition 9. The space

$$L_{p,q}(\Sigma) \equiv L_p(\Omega; L_q(\Gamma_x)) := \{u : \Sigma \to \mathbb{R} \text{ measurable such that } u(x) \in L_q(\Gamma_x) \text{ for a.e. } x \in \Omega \text{ and } \|u\|_{L_p(\Omega; L_q(\Gamma_x))} < \infty\}$$

is a Banach space with the norm

$$\|u\|_{L_{p,q}(\Sigma)} := \begin{cases} (\int_\Omega \|u(x)\|^p_{L_q(\Gamma_x)}\, dx)^{1/p}, & p < \infty, \\ \operatorname{ess\,sup}_{x\in\Omega} \|u(x)\|_{L_q(\Gamma_x)}, & p = \infty. \end{cases}$$

and a Hilbert space in case of $p = q = 2$ with the obvious scalar product.

Proof. It is clear by construction that $u(x, \cdot) : \Gamma_x \to \mathbb{R}$ is a measurable function and the norm $\|u(x, \cdot)\|_{L_q(\Gamma_x)}$ is measurable in x. Hence, the space above is well-defined. The proof of completeness is obtained by slight modification of the usual Fischer-Riesz type arguments. See [16] for details. □

Proposition 10. Let $p < \infty$. Then the space $L_{p,p}(\Sigma)$ is equivalent to $L_p(\Sigma)$. In general, both spaces are *not* isometric isomorph.

Proof. Let $f \in L_{p,p}(\Sigma)$. Then, by Fubini's theorem, the mapping $x \mapsto \int_{\Gamma_x} f(x, y)\, d\sigma_y$ is measurable and

(9)
$$\begin{aligned} \|f\|^p_{L_{p,p}(\Sigma)} &= \int_\Omega \int_{\Gamma_x} |f(x, y)|^p\, d\sigma_y\, dx \\ &= \sum_{k=1}^N \int_\Omega \int_{W^{m-1}} (\eta^k |f|^p)\left(x, (\Psi^k_x)^{-1}(\zeta', 0)\right) \sqrt{1 + |\nabla_{\zeta'} g^k_x(\zeta')|^2}\, d\zeta'\, dx. \end{aligned}$$

Moreover, by construction, the integral of f over Σ can be written as

(10)
$$\begin{aligned} \|f\|^p_{L_p(\Sigma)} &= \int_\Sigma |f(x, y)|^p\, d\sigma_{x,y} \\ &= \sum_{k=1}^N \int_{W^{n+m-1}} (\eta^k |f|^p)\left(\Psi^k)^{-1}(\xi', 0)\right) \sqrt{1 + |\nabla_{\xi'} g^k(\xi')|^2}\, d\xi' \\ &= \sum_{k=1}^N \int_\Omega \int_{W^{m-1}} (\eta^k |f|^p)\left(x, (\Psi^k_x)^{-1}(\zeta', 0)\right) \sqrt{1 + |\nabla_x g^k(x, \zeta')|^2 + |\nabla_{\zeta'} g^k_x(\zeta')|^2}\, d\zeta'\, dx. \end{aligned}$$

The norm equivalence follows now from the fact that the surface elements in (9) and (10), namely

$$\sqrt{1 + |\nabla_{\zeta'} g^k_x(\zeta')|^2} \qquad \text{and} \qquad \sqrt{1 + |\nabla_x g^k(x, \zeta')|^2 + |\nabla_{\zeta'} g^k_x(\zeta')|^2},$$

are essentially bounded from above and below, uniformly w.r.t. x.

A counterexample that the spaces are not isometric isomorph is given as follows: Let $n = m = 1$, $\Omega = (a, b)$, $b > a > 0$, and $Y_x := (-x, x)$. If we integrate 1 over Γ_x, we obtain 2 in each cell. Therefore, the integration (9) gives $\int_\Omega \int_{\Gamma_x} d\sigma_y\, dx = 2(b-a)$ whereas an integration over the one-dimensional surface measure according to (10) gives the value $2(b - a)\sqrt{2}$.

□

Next we construct the connection between the spaces $W^{0,1}_{p,q}$ and $L_{p,q}(\Sigma)$ via the *distributed trace*.

Proposition 11. Let $p < \infty$ and $\gamma_x : W^1_q(Y_x) \to L_q(\Gamma_x)$ be the continuous trace operator. Then $\|\gamma_x\|$ is uniformly bounded in $x \in \Omega$ and the distributed trace

$$\gamma : W^{0,1}_{p,q}(Q) \to L_{p,q}(\Sigma), \qquad \gamma(u)(x) = \gamma_x(u(x)),$$

is a bounded linear operator.

Proof. By construction, γ_x is measurable and the norm $\|\gamma_x\|$ is bounded uniformly w.r.t. x. (See [16] for details and for an estimation of $\|\gamma_x\|$ in terms of the parametrisation.) The boundedness of the distributed trace follows then from the estimate

$$\|\gamma u\|^p_{L_{p,q}(\Sigma)} = \int_\Omega \|\gamma_x u(x)\|^p_{L_q(\Gamma_x)}\, dx \le \sup_{x\in\Omega} \|\gamma_x\|^p \cdot \int_\Omega \|u(x)\|^q_{W^1_q(Y_x)}\, dx.$$

□

Remark. Higher order trace estimates in $W^l_q(Y_x)$, $l > 1$, can also be formulated. The only technical difficulty is the higher regularity needed for the boundary Γ_x. The parametrisation has to be adapted to the case where the geometry is smoother with respect to y than w.r.t. x. We do not follow this direction.

4. Application to reactive transport in porous media

Making use of the space constructions in the previous sections, we can now define a variational formulation of a distributed-microstructure system for reactive transport in a heterogeneous porous medium and prove its wellposedness.

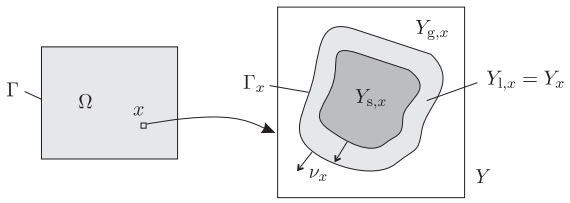


FIGURE 3. Geometry of the distributed microstructure model (11).

4.1. The problem. We consider an unsaturated porous medium, in which the distribution of pore water is assumed completely known and transport of water is at rest. Let $S = [0, T]$ be a bounded time interval and $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain with boundary $\Gamma = \partial\Omega$, representing the macroscopic domain filled by the medium.

Let $Y := (0, l)^n$ denote the n-dimensional cuboid with side length $l > 0$. For each $x \in \Omega$, assume we are given bounded domains $Y_{\mathrm{s},x}, Y_x \subset Y$ representing the solid and the liquid phase near x such that $Y_{\mathrm{s,x}} \cap Y_x = \emptyset$. We assume that the pore air is connected, whereas the liquid domains Y_x are individually isolated, i.e., $\overline{Y}_x \subset Y$. An example of such a geometry is depicted in figure 3. In typical applications, this corresponds to a *low* humidity of the medium. We then denote the interface between the gaseous and the liquid phase by $\Gamma_x := \partial Y_x \setminus \partial Y_{\mathrm{s},x} \neq \emptyset$ and assume that the family $\{(Y_x, \Gamma_x) : x \in \Omega\}$ is a regular distribution of cells in the sense of Def. 7.

We consider a substance A that diffuses *slowly* as a solute in the pore water and *fast* as a gas in the pore air. At the gas–liquid interfaces, exchange of A occurs in both directions. In the pore water, A is subjected to one or more chemical reactions. The setting is motivated by carbonation of concrete; see [24, 18]. The mass balance for the concentration $u = u(t,x)$ of A in the pore air is effectively described by an averaged diffusion equation of the form

$$\partial_t(\theta(x)u(t,x)) - \operatorname{div}(d(x)\nabla u) = -f(t,x), \qquad t \in S,\ x \in \Omega, \tag{11a}$$

where $\theta(x) = |Y_{\mathrm{g},x}|/|Y|$ is the volume fraction of the pore air, $d(x) \in \mathbb{R}^{n\times n}$ is the effective diffusion tensor and $f(t,x)$ is the amount of A that gets absorbed in the pore water. The structure of f will be derived below. At $\Gamma = \partial\Omega$ we impose the Robin condition

$$-d\nabla u(t,x)\cdot\nu(x) = b(x)(u - u^{\mathrm{e}}(t,x))), \qquad t \in S,\ x \in \Gamma, \tag{11b}$$

with the unit normal $\nu(x)$ (directed outward), an exchange coefficient $b(x)$ and a given external concentration $u^{\mathrm{e}}(t,x)$.

The slower transport of A in the pore water is described at the micro scale. Let $U = U(t,x,y)$ be the concentration of A at time $t \in S$, macroscopic coordinate $x \in \Omega$ and microscopic coordinate $y \in Y_x$. Then the microscopic (not-averaged!) mass balance reads

$$\partial_t U(t,x,y) - \operatorname{div}_y(D\nabla_y U) = g(t,x,y,U), \qquad t \in S,\ x \in \Omega,\ y \in Y_x, \tag{11c}$$

where $D > 0$ is the microscopic diffusivity and $g(t,x,y,U)$ is the production or consumption of A by chemical reactions. Note that all spatial derivatives in (11c) are taken with respect to the microscopic coordinate y, such that x is effectively a parameter in (11c). We indicate this by the lower index y. In contrast, the symbols "∇" and "div" without lower index stand for differentiation w.r.t. x.

For a function $w : Y_x \to \mathbb{R}$, let $\gamma_x w := w|_{\Gamma_x}$ be its trace at the boundary Γ_x. Then the boundary conditions for U read

$$-D\nabla_y U(t,x,y)\cdot\nu_x = k(\gamma_x U(t,x,y) - Hu(t,x)), \qquad t \in S,\ x \in \Omega,\ y \in \Gamma_x, \tag{11d}$$

where ν_x is the unit normal to ∂Y_x (directed outward), k is an exchange coefficient and H is the equilibrium constant between u and U (*Henry constant*), and

$$-D\nabla_y U(t,x,y)\cdot\nu_x = 0 \qquad t \in S,\ x \in \Omega,\ y \in \partial Y_x \setminus \Gamma_x. \tag{11e}$$

Now the *total amount of A crossing the interface* Γ_x at a given time $t \in S$ and a given point $x \in \Omega$ is

$$f(t,x) = \frac{1}{|Y|}\int_{\Gamma_x} k(Hu - \gamma_x U)\,d\sigma_y, \qquad t \in S,\ x \in \Omega, \tag{11f}$$

which completes the mass balance (11a) for u. Finally, the initial conditions are

$$u(0,x) = u_0(x), \qquad U(0,x,y) = U_0(x,y), \qquad x \in \Omega,\ y \in Y_x. \tag{11g}$$

The system (11) is a semilinear, weakly coupled system of two parabolic PDEs. It is conceptually similar to the *regularised microstructure model* introduced in [25] for flow in fissured media.

4.2. Variational formulation. Let $\theta \in L_\infty(\Omega)$ be bounded away from zero, i.e. $\theta(x) \geq \theta_0 > 0$ for a.e. $x \in \Omega$. Let $d : \Omega \to \mathbb{R}^{n\times n}$ be measurable and uniformly elliptic, i.e. there exist constants $c, C > 0$ such that

$$c\,|\xi|^2 \leq \sum_{k,l} d_{k,l}(x)\xi_k\xi_l \leq C\,|\xi|^2 \qquad \forall \xi \in \mathbb{R}^n, \text{ a.e. } x \in \Omega.$$

Let $b \in L^\infty(\partial\Omega)$ be nonnegative and let k, H and D be positive constants. For the initial and external values we assume that $u_0 \in L^\infty(\Omega)$, $U_0 \in L^\infty(\Omega \times Y_x)$ and $u^{\mathrm{e}} \in L^\infty(S \times \partial\Omega)$ such that for a positive constant C_U it holds

$$0 \leq Hu_0(x), U_0(x,y), Hu^{\mathrm{e}}(t,x) \leq C_U \quad \text{a.e.} \tag{12}$$

The reaction term has the structure

$$g(t,x,y,U) = -M_1(t,x,y)\eta_1(U) + M_2(t,x,y)\eta_2(U), \tag{13}$$

where $M_1, M_2 \in L_\infty(S \times \Omega \times Y_x)$ are nonnegative functions and $\eta_1, \eta_2 : \mathbb{R} \to [0,\infty)$ are locally Lipschitz and satisfy

$$\eta_1(U) = 0 \quad \text{for } U \leq 0 \qquad \text{and} \qquad \eta_2(U) \leq 1 + |U| \quad \text{for } U \in \mathbb{R}. \tag{14}$$

Note that (14) actually implies that the production rate η_2 is *globally* Lipschitz on $[0,\infty)$.

We denote by H_θ the usual space $L_2(\Omega)$, equipped with the equivalent scalar product

$$(u,\, v)_{H_\theta} := \int_\Omega \theta(x)u(x)v(x)\,dx, \qquad u, v \in H_\theta,$$

and introduce

$$\begin{aligned} V &:= H^1(\Omega) \times L_2(\Omega; H^1(Y_x)),\\ H &:= H_\theta \times L_2(\Omega; L_2(Y_x)),\\ \mathcal{V} &:= L_2(S; V), \quad \mathcal{V}' := L_2(S; V'). \end{aligned}$$

Then, by Def. 4 and Prop. 5, V and H are separable Hilbert spaces and the embeddings $V \hookrightarrow H \hookrightarrow V'$ are continuous and dense. Moreover, if $\gamma_x : H^1(Y_x) \to L_2(\Gamma_x)$ is the usual trace map on the cell boundary, then from Prop. 11 we obtain that the *distributed trace* $\gamma : L_2(\Omega; H^1(Y_x)) \to L_2(\Omega; L_2(\Gamma_x))$ is a bounded linear operator. We introduce the notation

$$(u,\, v)_{\Omega\times\Gamma_x} := \int_\Omega\int_{\Gamma_x} uv\,d\sigma_y dx, \qquad u, v \in L_2(\Omega; L_2(\Gamma_x)), \tag{15}$$

and $(\cdot,\,\cdot)_{\Omega\times Y_x}$, $(\cdot,\,\cdot)_\Omega$, etc., analogously. Note that (15) is *not* equal to the integral over the $(2n-1)$-dimensional manifold $\Omega \times \Gamma_x$ (cf. Prop. 10).

Definition 12 (Weak upper and lower solutions)**.** A pair of essentially bounded functions $[u, U] \in \mathcal{V}$ is called a weak lower (upper) solution of problem (11) if $[u(0), U(0)] \leq (\geq)\, [u_0, U_0]$, and for all $[\varphi, \Psi] \in V$ with $[\varphi, \Psi] \geq 0$ a.e., it holds

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\,([u,U],\,[\varphi,\Psi])_H + (d\nabla u,\,\nabla\varphi)_\Omega + (D\nabla_y U,\,\nabla_y\Psi)_{\Omega\times Y_x} + (b(u-u^{\mathrm{e}}),\,\varphi)_\Gamma\\ &\quad + \big(k(Hu-\gamma U),\,|Y|^{-1}\varphi - \gamma\Psi\big)_{\Omega\times\Gamma_x} \leq (\geq)\ (g(\cdot,U),\,\Psi)_{\Omega\times Y_x} \qquad \text{for a.e. } t \in S. \end{aligned} \tag{16}$$

If $[u, U]$ is both a lower and an upper weak solution, then it is called a weak solution.

Remark. The system for u, U is *quasi-monotone increasing* in the sense of [20]. We modify the technique of weak upper and lower solutions and the comparison principle from [14, 9].

Proposition 13 (Comparison Principle)**.** Assume that g is globally Lipschitz in U. Let $[\underline{u}, \underline{U}]$ and $[\overline{u}, \overline{U}]$ be lower and upper weak solutions, resp., corresponding to different data satisfying $\underline{u}_0, \leq \overline{u}_0$, $\underline{U}_0 \leq \overline{U}_0$ and $\underline{u}^{\mathrm{e}} \leq \overline{u}^{\mathrm{e}}$ a.e. Then

$$\underline{u}(t,x) \leq \overline{u}(t,x),\ \underline{U}(t,x,y) \leq \overline{U}(t,x,y) \qquad \text{for a.e. } t \in S,\ x \in \Omega,\ y \in Y_x.$$

Proof. Let $\beta = (H|Y|)^{-1}$ and denote $u = \underline{u} - \overline{u}$, $u_0 = \underline{u}_0 - \overline{u}_0$, etc. Subtracting both inequalities (16) for $[\underline{u}, \underline{U}]$ and for $[\overline{u}, \overline{U}]$ and testing with

$$[\varphi, \Psi](t) = [u^+(t), \beta U^+(t)], \qquad t \in S,$$

gives with standard arguments for $\tau \in (0, T]$

$$\begin{aligned}\frac{1}{2}\|[u^+(\tau), \beta U^+(\tau)]\|_H^2 &+ \int_0^\tau \big(d\nabla u,\, \nabla u^+\big)_\Omega + \int_0^\tau \big(D\nabla_y U,\, \beta \nabla_y U^+\big)_{\Omega\times Y_x} \\ &+ \int_0^\tau \big(bu,\, u^+\big)_\Gamma + \int_0^\tau \big(k(Hu - \gamma U),\, \beta(Hu^+ - \gamma U^+)\big)_{\Omega\times\Gamma_x} \\ &\leq \frac{1}{2}\|[u_0^+, \beta U_0^+]\|_H^2 + \int_0^\tau \big(bu^{\mathrm{e}},\, u^+\big)_\Gamma + \int_0^\tau \big(g(\cdot, \underline{U}) - g(\cdot, \overline{U}),\, \beta U^+\big)_{\Omega\times Y_x}.\end{aligned}$$

Since the cut-off function $(\cdot)^+ : \mathbb{R} \to \mathbb{R}^{\geq 0}$ is monotone, it holds

$$\int_0^\tau \big(k(Hu - \gamma U),\, \beta(Hu^+ - \gamma U^+)\big)_{\Omega\times\Gamma_x} \geq 0.$$

Since g is globally Lipschitz in U, it follows that

$$\begin{aligned}\frac{1}{2}\|[u^+(\tau), \beta U^+(\tau)]\|_H^2 &+ c\int_0^\tau \|\nabla u^+\|_\Omega^2 + A\beta \int_0^\tau \|\nabla_y U^+\|_{\Omega\times Y_x}^2 \\ &\leq \frac{1}{2}\|[u_0^+, \beta U_0^+]\|_H^2 + \frac{1}{2}\int_0^\tau \|\sqrt{b}(u^{\mathrm{e}})^+\|_\Gamma^2 + C\int_0^\tau \|U^+\|_{\Omega\times Y_x}^2.\end{aligned}$$

By Gronwall's inequality one obtains

$$\begin{aligned}\|[(\underline{u} - \overline{u})^+, (\underline{U} - \overline{U})^+]\|_{L_\infty(S;H)} &+ \|[(\underline{u} - \overline{u})^+, (\underline{U} - \overline{U})^+]\|_{L_2(S;V)} \\ &\leq C\Big(\|(\underline{u}^{\mathrm{e}} - \overline{u}^{\mathrm{e}})^+\|_{L_2(S\times\Gamma)} + \|[(\underline{u}_0 - \overline{u}_0)^+, (\underline{U}_0 - \overline{U}_0)^+]\|_H\Big).\end{aligned}$$

Now the result follows immediately. □

By similar arguments, we obtain an energy estimate for the system.

Proposition 14 (Energy estimate)**.** There exists a constant $C > 0$ such that every weak solution $[u, U] \in \mathcal{V}$ satisfies

$$\|[u,U]\|_{L_\infty(S;H)} + \|[u,U]\|_{L_2(S;V)} \leq \Big(C(1 + \|u^{\mathrm{e}}\|_{L_2(S\times\Gamma)} + \|[u_0, U_0]\|_H\Big) \tag{17}$$

and

$$\|[u', U']\|_{L_2(S;V')} \leq C\Big(\|[u,U]\|_{L_2(S;V)} + \|u^{\mathrm{e}}\|_{L_2(S\times\Gamma)} + \|g(\cdot, U)\|_{L_2(S\times\Omega\times Y_x)}\Big). \tag{18}$$

Remark. Due to assumption (14), global Lipschitz continuity of g is *not* needed for the result. However, if either g is globally Lipschitz in U or if an a-priori L_∞-bound for U is known, then the *a-posteriori* estimate (18) for the time derivative can be turned into an *a-priori* estimate using (17).

4.3. Boundedness, existence and uniqueness. First, we are looking for candidates for upper and lower solutions.

Proposition 15 (**Positivity and boundedness**)**.** Let $\overline{U} \in C^1(S)$ be a solution of the ODE

$$\partial_t \overline{U} = \|M_2\|_\infty \eta_2(\overline{U}), \qquad t \in S,$$

satisfying $\overline{U}(0) \geq C_U$ where C_U and M_2 are given from (12) and (13). Then each solution $[u, U]$ satisfies

$$0 \leq u(t,x) \leq H^{-1}\overline{U}(t), \quad 0 \leq U(t,x,y) \leq \overline{U}(t) \qquad \text{for a.e. } t \in S,\ x \in \Omega,\ y \in Y_x.$$

Proof. Let $[u, U]$ be a solution. Since by definition $[u, U]$ is essentially bounded, we can replace η_1 w.l.o.g by a cut-off function. So g can be assumed *globally* Lipschitz in U. Hence, Prop. 13 can be applied. Now the nonnegativity result follows from the fact that by (14) $g(\cdot, 0) \geq 0$ and therefore $[0, 0]$ is a weak lower solution.

Denote $\overline{u} := H^{-1}\overline{U}$. Then one has to check that $[\overline{u}, \overline{U}]$ is a weak upper solution: By (12) we have

$$[\overline{u}(0), \overline{U}(0)] = [H^{-1}\overline{U}(0), \overline{U}(0)] \geq [u_0, U_0]$$

and also $b(\overline{u} - u^{\mathrm{e}}) \geq 0$ a.e. Finally, for nonnegative test functions $[\varphi, \Psi] \in V$, we have

$$\begin{aligned} \big([\overline{u}', \overline{U}'], [\varphi, \Psi]\big)_H &= \big(\theta\overline{u}', \varphi\big)_\Omega + \big(\overline{U}', \Psi\big)_{\Omega \times Y_x} \\ &\geq 0 + \big(\|M_2\|_\infty \eta_2(U), \Psi\big)_{\Omega \times Y_x} \\ &\geq \big(g(U), \Psi\big)_{\Omega \times Y_x}. \end{aligned}$$

This gives the inequality (16) for $[\overline{u}, \overline{U}]$. □

Remark. Note that the cutoff argument works only for the *negative* reaction term η_1. The crucial point is that, after cutting off the function η_1 for values above $M > 0$, say, we are able to prove L_∞-bounds for U that are *independent* of M. This is not possible for the positive part η_2.

Theorem 16 (Existence and uniqueness)**.** There exists a unique weak solution of problem (11).

Proof. We use Banach's fixed point theorem with the weighted-norm space

$$X = C(S; H), \qquad \|u\|_X = \max_{t \in S} \{e^{-\lambda t} \|u(t)\|_H\}, \qquad \lambda > 0.$$

For $[\tilde{u}, \tilde{U}] \in X$ given, we consider the linearised problem (16) for $[u, U]$, in which the right-hand side is replaced by $\Big(g(\cdot, \tilde{U}), \Psi\Big)_{\Omega \times Y_x}$. By Prop. 14, we can w.l.o.g. assume g to be globally Lipschitz. By a standard result on evolution equations (see, e.g., [23], Thm. 10.3), there exists a unique solution $[u, U] \in L_2(S; V) \cap H^1(S; V') \hookrightarrow X$.

Now we define a fixed point operator as

$$T : X \to X, \qquad T([\tilde{u}, \tilde{U}]) = [u, U],$$

and consider solutions $[u, U]$, $[v, V]$ corresponding to different data $[\tilde{u}, \tilde{U}]$ and $[\tilde{v}, \tilde{V}]$. By similar arguments as in Prop. 13, we obtain, for $\tau \in (0, T]$, the estimate

$$\begin{aligned}\|[u(\tau) - v(\tau), U(\tau) - V(\tau)]\|_H^2 &\leq C \int_0^\tau \|[\tilde{u}(t) - \tilde{v}(t), \tilde{U}(t) - \tilde{V}(t)]\|_H^2 \, dt \\ &\leq C \int_0^\tau e^{\lambda t} \, dt \max_{t \in [0,\tau]} \left\{ e^{-\lambda t} \|[\tilde{u}(t) - \tilde{v}(t), \tilde{U}(t) - \tilde{V}(t)]\|_H^2 \right\} \\ &\leq C \frac{e^{\lambda \tau}}{\lambda} \|[\tilde{u} - \tilde{v}, \tilde{U} - \tilde{V}]\|_X^2.\end{aligned}$$

It follows that

$$\|[u - v, U - V]\|_X \leq \sqrt{\frac{C}{\lambda}} \|[\tilde{u} - \tilde{v}, \tilde{U} - \tilde{V}]\|_X,$$

and, hence, for λ chosen small enough, $T : X \to X$ is strictly contractive. □

Remark.

(1) Note that the embedding $V \hookrightarrow H$ is *not compact*. For this reason, we have chosen a technique of proving existence that does not need any compactness arguments.

(2) Uniqueness can alternatively be obtained by the analogous estimate as in the proof of Prop. 13, applied to two solutions and omitting the cut-off functions. This procedure yields also continuous dependence of the solutions on the initial and boundary data.

Acknowledgments

The authors would like to express their gratitude to the state of Bremen for funding this work via the PhD program *Scientific Computing in Engineering (SCiE)* as well as to the German National Science Foundation (DFG) for a grant through the special priority program 1122 *Prediction of the Course of Physicochemical Damage Processes Involving Mineral Materials*.

References

[1] T. Arbogast. Analysis of the simulation of single phase flow through a naturally fractured reservoir. *SIAM J. Numer. Anal.*, 26(1):12–29, 1989.

[2] T. Arbogast, J. Douglas, Jr., and U. Hornung. Deriving the double porosity model of single phase flow via homogenization theory. *SIAM J. Math. Anal.*, 21:823–863, 1990.

[3] T. Arbogast, J. Douglas, Jr., and U. Hornung. Modeling of naturally fractured reservoirs by formal homogenization techniques. In R. Dautray, editor, *Frontiers in Pure and Applied Mathematics*, pages 1–19. Elsevier, 1991.

[4] A. Benedek and R. Panzone. The space L^p, with mixed norm. *Duke Math. J.*, 28:301–324, 1961.

[5] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory*. Springer, 1998.

[6] R. Dautray and J.-L. Lions. *Spectral theory and applications*, volume 3 of *Mathematical analysis and numerical methods for science and technology*. Springer, 1992.

[7] J. Dixmier. *Von Neumann algebras*. North-Holland, 1981.

[8] N. Dunford and J. T. Schwartz. *Linear operators I*. Interscience Publishers, 1971.

[9] A. Friedman and P. Knabner. A transport model with micro- and macrostructure. *J. Differ. Equations*, 98(2):328–354, 1992.

[10] A. Friedman and A. T. Tzavaras. A quasilinear parabolic system arising in modeling of catalytic reactors. *J. Differ. Equations*, 70:167–196, 1987.

[11] U. Hornung, editor. *Homogenization and Porous Media*. Springer, 1997.

[12] U. Hornung and W. Jäger. Diffusion, convection, adsorption, and reactions of chemicals in porous media. *J. Differ. Equations*, 92:199–225, 1991.

[13] U. Hornung, W. Jäger, and A. Mikelić. Reactive transport through an array of cells with semi-permeable membranes. *RAIRO Modél. Math. Anal. Numér.*, 28(1):59–94, 1994.

[14] P. Knabner. *Mathematische Modelle für Transport und Sorption gelöster Stoffe in porösen Medien*. Verlag Peter Lang, 1991.
[15] A. Kufner, O. John, and S. Fuçík. *Function Spaces*. Academia, Publishing house of the Czecholovak Academy of Sciences, 1977.
[16] S. A. Meier. *Two-scale models of reactive transport and evolving microstructure*. PhD thesis, University of Bremen, 2008.
[17] S. A. Meier and M. Böhm. On a micro-macro system arising in diffusion–reaction problems in porous media. In M. Fila et al., editors, *Proceedings of the Equadiff 11 (Bratislava, 2005)*, pages 259–263, 2005.
[18] S. A. Meier, M. A. Peter, and M. Böhm. A two-scale modelling approach to reaction-diffusion processes in porous materials. *Comp. Mat. Sci.*, 39:29–34, 2007.
[19] M. Milla Miranda and J. Límaco Ferrel. The Navier-Stokes equation in noncylindrical domain. *Mat. Apl. Comput.*, 16(3):247–265, 1997.
[20] C. V. Pao. *Nonlinear Parabolic and Elliptic Equations*. Plenum Press, 1992.
[21] M. Peszyńska. *Flow through fissured media. Mathematical analysis and numerical approach*. PhD thesis, University of Augsburg, 1992.
[22] M. A. Peter. Homogenisation in domains with evolving microstructure. *C. R. Mécanique*, 335(7):357–362, 2007.
[23] M. Renardy and R. C. Rogers. *An introduction to partial differential equations*. Springer, 1996.
[24] A. V. Saetta, B. A. Schrefler, and R. V. Vitaliani. The carbonation of concrete and the mechanism of moisture, heat and carbon dioxide flow through porous materials. *Cem. Conc. Res.*, 23(4):761–772, 1993.
[25] R. E. Showalter and N. J. Walkington. Micro-structure models of diffusion in fissured media. *J. Math. Anal. Appl.*, 155:1–20, 1991.
[26] M. Takesaki. *Theory of operator algebras*. Springer, 2003.
[27] J. Wloka. *Partielle Differentialgleichungen*. Teubner, 1982.
[28] K. Yosida. *Functional analysis*. Springer, 1978.

Centre for Industrial Mathematics, University of Bremen, Postfach 330 440, 28334 Bremen, Germany

E-mail: sebam@math.uni-bremen.de and mbohm@math.uni-bremen.de
URL: http://www.math.uni-bremen.de/zetem/modpde/indexmod.html

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 126–149

NUMERICAL METHODS FOR UNSATURATED FLOW WITH DYNAMIC CAPILLARY PRESSURE IN HETEROGENEOUS POROUS MEDIA

MAŁGORZATA PESZYŃSKA AND SON-YOUNG YI

Abstract. Traditional unsaturated flow models use a capillary pressure-saturation relationship determined under static conditions. Recently it was proposed to extend this relationship to include dynamic effects and in particular flow rates. In this paper, we consider numerical modeling of unsaturated flow models incorporating dynamic capillary pressure terms. The resulting model equations are of nonlinear degenerate pseudo-parabolic type with or without convection terms, and follow either Richards' equation or the full two-phase flow model. We systematically study the difficulties associated with numerical approximation of such equations using two classes of methods, a cell-centered finite difference method (FD) and a locally conservative Eulerian-Lagrangian method (LCELM) based on the finite difference method. We discuss convergence of the methods and extensions to heterogeneous porous media with different rock types. In convection-dominated cases and for large dynamic effects instabilities may arise for some of the methods while those are absent in other cases.

Key Words. unsaturated flow, Richards' equation, two-phase flow model, dynamic capillary pressure, pseudo-parabolic equation, finite difference method, locally conservative Eulerian-Lagrangian method, implicit time-stepping

1. Introduction

The main interest of this paper is in numerical algorithms for unsaturated flow in highly heterogeneous media and in particular handling dynamic capillary pressure.

Unsaturated preferential flow in porous media is a physical phenomenon occurring in heterogeneous soils and bedrock and is related to the presence of special features of the medium such as cracks, fissures, and macropores. Such heterogeneities are represented in partial differential equation (PDE) models of the flow by a variation of nonlinear *rock properties* of the medium with position, called the *rock type* dependence. Here we are concerned mainly with the capillary pressure function; that is, the pressure-saturation relationship $S \mapsto P_c(S)$ which, when this property is rock type dependent, it reads $S \mapsto P_c(\mathbf{x}, S)$, where $\mathbf{x}$ denotes position. It is standard practice to include rock type dependence in a reservoir simulator [47]; however, there are few associated mathematical and numerical analyses except [18, 33] and those for multiscale heterogeneities developed in [28, 19, 16, 17].

Additional phenomena occurring in preferential flow such as nonequilibrium effects, hysteresis, and/or large flow velocities have been recently discussed by experimental and theoretical soil physicists. In particular, it has been observed and reconfirmed recently, see [66] and references therein, that rock properties measured

1991 *Mathematics Subject Classification.* AMS(MOS) subject classifications. 35K65, 35K70, 65M06, 76S05.

in a laboratory in equilibrium conditions bear little resemblance to those observed in experiments in case of large fluxes (velocities); it has been in fact postulated that the data collected especially for the capillary pressure has a non-unique character when considered over a range of nonequilibrium conditions. This suggests existence of a hidden variable as pointed out in [41] and can be explained in terms of the imbibition (increasing S) and drainage (decreasing S) hysteresis. Another recently proposed class of model modifications has been the use of *dynamic capillary pressure* [42, 68, 67, 66, 12, 11] which accounts for the dynamic flow rates via $S \mapsto P_c(\mathbf{x}, S, \frac{dS}{dt})$. Some authors believe that dynamic capillary pressure terms could explain instabilities in gravity-driven flow and in particular the phenomena of fingering [51]. Finally, there is evidence that in the presence of strong heterogeneities the conditions at the interface of different rock types should reflect non-equilibrium [55, 19].

Our interest is in numerical modeling of preferential flow in porous media which may occur at more than one scale through macropores or due to small or large inhomogeneities as well as in rock fractures, gravel filled excavated areas etc. Therefore, it is necessary for us to explore the numerical algorithms for dynamic capillary pressure and multiple rock types.

From a theoretical PDE point of view, Richards' model of unsaturated flow is a nonlinear degenerate parabolic equation in the unknown water saturation S; its character depends on the nonlinear diffusion parameter $\mathbf{D}(S)$ which may become degenerate (zero or very large) for some values of S. In addition, if the flow has vertical components, then the associated nontrivial convective term competes with the nonlinear degenerate diffusion. Depending on the rock type and initial and boundary conditions of the flow, the solutions may be smooth or may exhibit sharp fronts [4, 5, 6].

The presence of dynamic capillary pressure terms changes the type of original nonlinear degenerate parabolic PDEs to *pseudo-parabolic*, with the additional nonlinear degenerate term being proportional to a coefficient τ, see development in Section 2.3. Available existence, uniqueness, and regularity theory for pseudo-parabolic equations [58, 60, 61] predicts that the additional pseudo-parabolic term decreases the smoothing property characteristic to parabolic problems (if at all present) to a factor involving $e^{-\tau}$. In addition it is known that there is in general no maximum principle for pseudo-parabolic equations such as one expected of solutions to parabolic equations. Finally, we note that the available theory may or may not include cases with dominating and degenerate convection; assumptions need to be verified on a case by case basis.

Numerical methods for Richards' equation include practical implementations of finite difference [69], finite elements [69, 40, 44, 54], finite volumes [1] and characteristic-based methods [7]; we do not attempt to give a comprehensive review here. Typically, convergence results are formulated for either transformed variables or for cases away from degeneracy, or for regularized problem. See, e.g., [9, 36] for results and references using Kirchoff transformation, and [33] for those using similarity solutions. Numerical methods for Richards' equation in physical variables have been used in hydrology [20] but have not been analyzed outside smooth regimes where standard convergence rates apply. Furthermore, there exist a plethora of methods applicable to two-phase flow formulation of unsaturated flow, see [25, 43, 22, 23, 21]; however, similarly to the case of Richards' equation, those results have been formulated for transformed variables or for smooth regime(s). Finally, comparison of two-phase flow versus Richards' formulations have been studied

in [62, 44, 64]; these point out that in some regimes the use of Richards' equation leads to loss of information.

On the other hand, there exist discretizations and associated analysis for (single) pseudo-parabolic equations given in [35, 10, 38, 39, 37] but these theoretical convergence results do not indicate any difficulties associated with large convective (advective) terms. Possibly such cases are ruled out by assumptions on the smoothness of the analytical solution necessary to obtain convergence results.

However, the reported numerical simulation results point out difficulties in incorporating dynamic capillary pressure and stipulate that the problems are properties of the numerical method [42] or of the PDE's themselves [51]. Most of these results focus on the modeling aspects of dynamic capillary pressure and on the identification of the coefficient τ; τ appears not to be constant especially in heterogeneous media [42, 49]. We also note that the combination of dynamic capillary pressure with *play-type hysteresis*, for a nonlinear non-degenerate case of horizontal flow, is discussed in [13]. Among these results, [42] and [49] appear to use traditional discretizations of Richards' and two-phase formulations, respectively. The work [51] seeks similarity and traveling wave solutions. The approach in [13] leads to solving a time-lagged coupled system of two equations, the first of which is elliptic and is solved for pressure, while the second is an ODE for saturation. In the quoted results we did not find detailed formulation of numerical algorithm, or convergence analysis/studies.

As a point of departure for this paper, we consider numerical methods which apply to nontransformed, nonregularized model(s) of unsaturated preferential flow in porous media with strong heterogeneities and with locally large fluxes. We are interested in both Richards' and two-phase formulations. We systematically study the difficulties associated with the numerical approximation of dynamic capillary pressure terms. Since the reported numerical simulation results on dynamic capillary pressure using traditional methods appear to exhibit instabilities, our approach is to carefully formulate different variants, study their convergence and finally detect the presence of instabilities if any. In an effort to be inclusive, we include a large set of prototypes (variants) of the traditional methods from the finite difference (FD) family to determine whether a particular detail of discretization may be the factor that triggers the instabilities. The variants therefore include various ways of averaging, implicit or semi-implicit solutions, and different choices of primary unknowns. In addition to the FD family, we formulate and investigate a different class of models which enables a locally conservative handling of convection terms (LCELM family) combined with FD treatment of nonlinear diffusion. All these discretizations are proposed for Richards' equation.

Moreover, we discuss a numerical formulation for the full two-phase flow model which is more general than Richards' equation; we allow for different rock types and dynamic capillary pressure and compare solutions to those for Richards' equation. Our formulations and discussion are supported by convergence studies and the numerical simulation results are illustrated. Finally, we consider a representative cell of a multiscale heterogeneous medium; discussion of a full simulation of preferential flow with dynamic effects is outside the scope here; to our knowledge such a study has not been undertaken.

Below in Section 2 we discuss the physical models, in Section 3 we formulate the numerical methods, and in Section 4 we present convergence studies and simulation results. The paper closes with Conclusions and Acknowledgements.

2. Physical model

Consider a porous medium: an open bounded domain $\Omega \subset \mathbb{R}^d$, with $d = 1, 2, 3$, characterized by physical properties of porosity $\phi = \phi(\mathbf{x}) > 0$ and (absolute) permeability $\mathbf{K} = \mathbf{K}(\mathbf{x}) \in \mathbb{R}^{d\times d}$, the latter in general heterogeneous and anisotropic. Here we assume as it is usually done that $\mathbf{K}$ is uniformly positive definite. In addition, let $\mathbf{K}$ be diagonal, i.e. its variability is aligned with coordinate axis.

Furthermore, let us be given for convenience, in the given coordinate system, the value of depth of the porous medium $D(\mathbf{x})$ under the earth's surface. We set

$$i)\, D(\mathbf{x}) \equiv 0: \text{ horizontal flow case,} \tag{1}$$

$$ii)\, D(\mathbf{x}) = x: \text{ vertical infiltration problem.} \tag{2}$$

The governing equations for the flow of two immiscible fluids in this porous medium are given by the full two-phase flow model (3)–(4) or by the Richards' equation model to be defined below (18) which is a simplified version of the two-phase model. The models are standard [52, 24, 43, 50]. We briefly recall the formulation(s) for completeness.

The two phases are denoted by subscripts α whereby $\alpha = w$ refers to water (wetting phase) and $\alpha = n$ refers to the other (nonwetting) phase which may be air or a hydrocarbon phase/component. The fluids have densities ρ_w, ρ_n and viscosities μ_w, μ_n, respectively. For simplicity, if no subscript is used, this denotes by default the wetting phase, i.e.. $\mu \equiv \mu_w, \rho \equiv \rho_w$ etc. The flow is described by the equations

$$\phi \frac{\partial S}{\partial t} - \nabla \cdot \left(\frac{1}{\mu} \mathbf{K} k_w(\mathbf{x}, S)(\nabla P - \rho G \nabla D) \right) = 0,\ \mathbf{x} \in \Omega,\ t > 0, \tag{3}$$

$$\phi \frac{\partial S_n}{\partial t} - \nabla \cdot \left(\frac{1}{\mu_n} \mathbf{K} k_n(\mathbf{x}, S)(\nabla P_n - \rho_n G \nabla D) \right) = 0,\ \mathbf{x} \in \Omega,\ t > 0, \tag{4}$$

$$P_n - P = P_c(\mathbf{x}, S). \tag{5}$$

In this model we have incorporated conservation of mass and multiphase extension of Darcy's law for immiscible two-phase fluids, which for each phase α, in the absence of external sources, read, respectively, $\phi \frac{\partial \rho_\alpha S_\alpha}{\partial t} + \nabla \cdot (\rho_\alpha \mathbf{u}_\alpha) = 0$ and $\mathbf{u}_\alpha = -\frac{\mathbf{K} k_\alpha}{\mu_\alpha}(\nabla P_\alpha - \rho_\alpha G \nabla D)$. The incompressibility assumption allows for the elimination of the constant densities from the formulation everywhere except in the gravity terms $\rho_\alpha G \nabla D$.

The unknowns of the system are the saturations S, S_n related by

$$S + S_n \equiv 1, \tag{6}$$

and the pressures P, P_n. These unknowns are related to each other by (5), or by its extension(s) to be discussed.

Here the rock-fluid properties are the capillary pressure relationship (5) and relative permeabilities k_w, k_n which are functions of the wetting phase saturation S and as such take values in $[0, 1]$ and are, respectively, nondecreasing and nonincreasing. Their dependence on $\mathbf{x}$ reflects heterogeneity of rocks and in particular the fact that in a domain composed of different rock types, for example, containing coarse and fine sand, the properties k_w, k_n, P_c will be given by different functional relationships, see Figure 1.

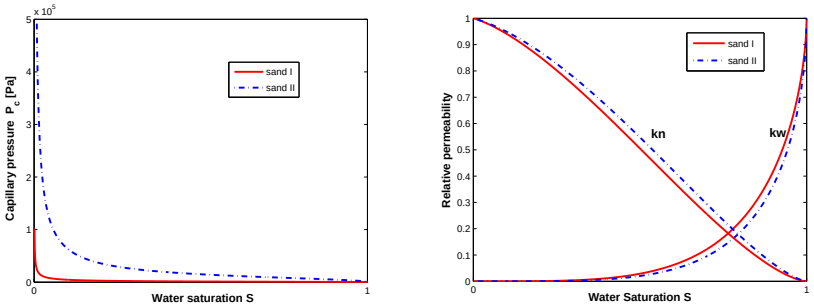


FIGURE 1. Capillary pressure (left) and relative permeabilities (right) for two rock types: sand I (coarse), and sand II (fine). Here we use van Genuchten model with data as in Table 4.1.

In numerical examples in this paper we use the so-called van-Genuchten–Mualem model in which

$$P_c(S) = \frac{1}{\alpha}(S^{-\frac{1}{m}} - 1)^{\frac{1}{n}}, \tag{7}$$

$$k_w(S) = S^{\epsilon}\left[1 - \left(1 - S^{\frac{1}{m}}\right)^{m}\right]^2, \tag{8}$$

(9)

$$k_n(S) = (1-S)^{\gamma}\left[1 - S^{\frac{1}{m}}\right]^{2m}. \tag{10}$$

Generally, $\epsilon = \frac{1}{2}$, $\gamma = \frac{1}{3}$ and $m = 1 - \frac{1}{n}$, and α, n are given from experiments. Rock type dependent k_w, k_n and P_c can be determined by $\alpha(\mathbf{x})$ and $n(\mathbf{x})$.

For the case when k_w, k_n, P_c are not rock type dependent, the existence and uniqueness as well as regularity of solutions to the PDEs have been studied. Given appropriate boundary and initial conditions and based on some assumptions on the data (see section below) one can determine the solution uniquely [4, 22, 23]. See also [18, 33] for the case of multiple rock types.

2.1. Boundary and initial conditions. Let us be given P_0, S_0 and P_D, S_D.

In the discussion below and in most experiments we use the following initial conditions

$$P(\mathbf{x}, 0) := const = P_0, \ \mathbf{x} \in \Omega, \tag{11}$$

$$S(\mathbf{x}, 0) := const = S_0, \ \mathbf{x} \in \Omega. \tag{12}$$

These conditions combined with (5) give initial values for the nonwetting phase pressure and saturation. Note that in the case (1) without gravity, a constant initial pressure and saturation represent an equilibrium solution. This is not true in case (2) with significant gravity where these initial conditions do not represent an equilibrium state. Finally, when considering multiple rock types in Example 3, we assume only (11) and an appropriately equivalent condition for the nonwetting phase, while $S(\mathbf{x}, 0)$ is determined from equality of capillary pressures.

The boundary $\partial\Omega$ consists of the no-flow part $\partial\Omega_N$ on which Neumann no-flow conditions are specified

$$\mathbf{u}_w \cdot \eta = 0, \ \mathbf{x} \in \partial\Omega_N, \ t > 0, \tag{13}$$

$$\mathbf{u}_n \cdot \eta = 0, \ \mathbf{x} \in \partial\Omega_N, \ t > 0 \tag{14}$$

as well as of its complement $\partial\Omega_D$ in $\partial\Omega$, the part on which we impose Dirichlet boundary conditions

$$P(\mathbf{x},t) = P_D,\ \mathbf{x} \in \partial\Omega_D, \tag{15}$$

$$S(\mathbf{x},t) = S_D,\ \mathbf{x} \in \partial\Omega_D. \tag{16}$$

In most experiments on part of $\partial\Omega_D$ we impose $P_D \equiv P_0$ and/or $S_D \equiv S_0$.

2.2. Richards' equation. Richards' equation can be derived from (3), (4), (5), (6) by assuming that the nonwetting phase, hereby presumed to be air, remains at a constant pressure equal to the atmospheric pressure which for convenience one can set (in certain units) to 0

$$P_n \equiv 0. \tag{17}$$

Thereby one of the equations and variables is eliminated and one rewrites the equation (3) as follows

$$\phi\frac{\partial S}{\partial t} - \nabla\cdot\left(\frac{1}{\mu}\mathbf{K}k_w(\mathbf{x},S)(\nabla P - \rho_w G\nabla D)\right) = 0, \tag{18}$$

where S, P are coupled via (5) and (17) as

$$P = -P_c(\mathbf{x},S). \tag{19}$$

If $P_c(\cdot)$ is a smooth invertible function, then one can simply seek a solution in either variable. However, in general, $P_c(\cdot)$ exhibits strong degenerate behavior so that $\lim_{S\to 0} P_c(S) = \infty$. In addition, in some range of S and for some rock types, $P_c'(S) \approx 0$. Therefore depending on the particular rock type and on specific issues with behavior of $P_c(\mathbf{x}, S)$, there are advantages in using either P or S as the primary unknown.

In hydrology applications [57, 50] yet another version of (18) is preferred. Here the (convective, first order) gravity term is separated from the diffusive second order term, the nonlinearites are lumped together, and it is assumed that $\mathbf{K} \equiv K\mathbf{I}$ and that the rock-fluid properties are not rock-dependent. Finally the chain rule is applied to set

$$\mathbf{D}(S) := -\frac{1}{\mu}\mathbf{K}k_w(S)P_c'(S), \tag{20}$$

$$\mathbf{C}(S) := \frac{1}{\mu}\mathbf{K}k_w(S)\rho_w G\nabla(-D(\mathbf{x})), \tag{21}$$

from which (18) can be rewritten as a convection diffusion equation solved for S

$$\phi\frac{\partial S}{\partial t} + \nabla\cdot(\mathbf{D}(S)\nabla S) = \nabla\cdot(\mathbf{C}(S)). \tag{22}$$

Here the *diffusivity* coefficient $\mathbf{D}(S)$ is nonnegative definite and degenerate and the advective term $\mathbf{C}(S)$ is monotone nonincreasing degenerate. In case (1) the advective term vanishes $\mathbf{C}(S) \equiv \mathbf{0}$ and the equation has a nonlinear degenerate parabolic character. In case (2) the problem has nontrivial advection which, depending on the magnitude of capillary pressure, may or may not dominate the character of the flow.

The equation (22) can be also reformulated in terms of other variables: in hydrology it is popular to use the water content $\Theta := \phi S$ and pressure head $h := \frac{P}{G\rho_w}$ [50, 57].

Yet another formulation and change of variable are used in derivation of existence, uniqueness, and regularity theory for Richards' equation. One identifies a smooth variable, say u, by a Kirchoff transformation for which the well-posedness

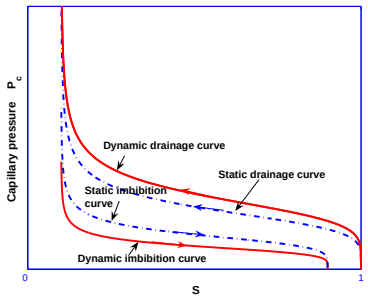


FIGURE 2. Idea of dynamic capillary pressure.

is studied; see [9, 36]. Then the values of S and/or P can be derived from u by means of a more or less degenerate transformation.

While the use of an auxiliary variable u is very helpful in understanding the transformed problem, it may or may not have a practical impact on applications in which the actual values of S and P need to be found. A similar remark applies to numerical methods applied to Richards' equation whose convergence for the transformed variable may be optimal in u albeit will exhibit sub-optimal convergence rates when studied in S, P.

2.2.1. Boundary and initial conditions. The boundary and initial conditions for Richards' equation follow from those for the two-phase flow (11), (12), (15), (13), (14), (16) noticing that by (17) the condition (14) is automatically satisfied and that (19) must be satisfied for all other conditions.

2.3. Dynamic capillary pressure. As mentioned above, the idea of incorporating dynamic effects in the capillary pressure-saturation relationship is to replace $P_c(\mathbf{x}, S)$ by $P_c(\mathbf{x}, S, \frac{\partial S}{\partial t})$ to account for dependence on time scale of getting to capillary equilibrium. Two main directions of models include the Hassanizadeh–model [42] and the Barenblatt–model [12, 11]. See Figure 2 for illustration. Also see a combination of play-type hysteresis and dynamic capillary pressure models in [13, 14].

In this paper we focus on the Hassanizadeh model and extend it to multiple rock types thereby assuming that (5) is replaced by (23)

$$P_c(\mathbf{x}, S, \frac{\partial S}{\partial t}) := P_c(\mathbf{x}, S) - \tau \frac{\partial S}{\partial t}, \tag{23}$$

where $\tau \geq 0$ is a constant or it varies with $\mathbf{x}, t$.

Note that plugging this relationship to (19) we obtain, instead of (18),

$$\phi \frac{\partial S}{\partial t} + \nabla \cdot \left(\frac{\mathbf{K} k_w(S)}{\mu} \nabla P_c(S, \frac{\partial S}{\partial t}) \right) = \nabla \cdot \left(\frac{\mathbf{K} k_w(S)}{\mu} \rho_w G \nabla D(x) \right) \tag{24}$$

which can be rewritten in a generic nonlinear pseudo–parabolic form similar to (22)

$$\frac{\partial S}{\partial t} - \nabla \cdot (\mathbf{D}(S) \nabla S) = \nabla \cdot \mathbf{C}(S) + \nabla \cdot \left(\frac{K k_w(S)}{\mu} \nabla \tau \frac{\partial S}{\partial t} \right). \tag{25}$$

2.4. Modeling flow in heterogeneous media. Consider now a full two-phase flow model with different rock types in a cell with coarse and fine sand described by $\mathbf{x} \in \Omega^I, \mathbf{x} \in \Omega^{II}$, respectively, in which fast and slow flow occur. In our experiments reported in Section 4.3 the parameters K^I and K^{II} differ by a factor of 10^3. See Figure 3 for a schematic representation of the cell.

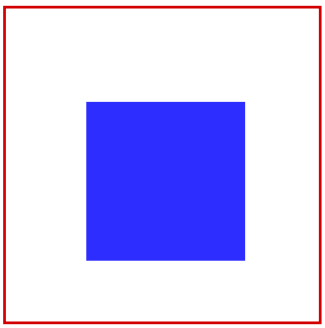


FIGURE 3. Schematic representation of a cell with two different media types: Ω^I (outside region) and Ω^{II} (inside).

It is well known [43] that in addition to variation in $\mathbf{K}$, one has to take into account rock-type dependent capillary-pressure properties

$$P_c(\mathbf{x}, S) = \begin{cases} P_c^I(S), & \mathbf{x} \in \Omega^I \\ P_c^{II}(S), & \mathbf{x} \in \Omega^{II} \end{cases}$$

as well as relative permeabilities k_w, k_n. In the dynamic capillary pressure model, we additionally have to take into account

$$\tau(\mathbf{x}) = \begin{cases} \tau^I, & \mathbf{x} \in \Omega^I \\ \tau^{II}, & \mathbf{x} \in \Omega^{II} \end{cases} .$$

3. Numerical approximation

Now we formulate two classes of numerical methods and their variants as applied to the two-phase problem system (3)–(6) and the Richards' equation in the form (18) or (22). These two methods are the cell–centered Finite Difference method (FD), and the Locally Conservative Eulerian-Lagrangian Method (LCELM). A subset of these methods handle multiple rock types; all algorithms handle dynamic capillary pressure terms directly; that is, via (23).

We refer to [56] for an analogy between mixed finite element spaces of lowest Raviart-Thomas order on rectangles and cell-centered FD methods for single-phase flow, and to [65] for a convergence analysis of FD for a linear elliptic problem. Theoretically, via equivalence to mixed finite elements, the cell-centered FD provide, for simple linear problems, convergence order in primary unknowns (P, S) similar to the one in fluxes, or higher, via superconvergence. However, handling of nonlinear and degenerate terms and convection dominated problems requires extensions to expanded mixed methods [9, 69] and higher order temporal discretizations with the additional error due to the loss of consistency order, or to the stabilizing terms. We refer to [52] for a standard discretization of multiphase flow and to [53] for extensions of [56] to multiphase flow and treatment of boundary conditions, and to [40] for a discussion of stabilization procedures. Also, see [9, 36, 1, 15] for a variety of other schemes. Again, we do not attempt to give a complete set of references.

Method	Model	advective	diffusive	time discr.	unknown
R_LCELM	Richards (22)	-	a	use $\overline{S}^n$. See (40)	S
R_LCELM2	Richards (22)	-	a	use S^{n-1}. See (40)	S
R_FD1	Richards (22)	u	a	time-lagging	S
R_FD2	Richards (18)	u	u	time-lagging	S
R_FD3	Richards (18)	u	u	fully implicit	P
R_FD4	Richards (18)	u	u	time-lagging	P
R_FD5	Richards (18)	u	u	fully implicit	S
2PH	Two-phase	u	u	fully implicit (variable t.step)	P, S
2PH_E	Two-phase	u	u	time-lagging	P, S
2PH_F	Two-phase	u	u	fully implicit (fixed t.step)	P, S

TABLE 1. Numerical approximation schemes considered in this paper. Symbol u denotes upwinding, symbol a denotes arithmetic average.

In our work FD is applied to both the two-phase formulation and Richards' equation; we discuss for simplicity only the discretization for $d = 1$ but the results shown come from a general MATLAB implementation applicable to $d = 1, 2, 3$. The variants of FD presented here differ by the way we handle a) edge nonlinearities, and b) time-discretization and solving the resulting nonlinear system. Only the two-phase FD method is applied when rock types are different. When upwinding is applied, we expect the method to be at most $O(\triangle x + \triangle T)$ accurate, in either variant, with restrictions on the time step for the non-implicit variants due to the CFL condition. In the case i) with no gravity and away from degeneracy of P_c, and when diffusion terms are discretized using arithmetic averaging which gives higher order consistency, one could expect a convergence rate of $O((\triangle x)^2 + \triangle T)$.

It is well known that standard finite difference and finite element methods may inaccurately approximate convection-diffusion problems when the Péclet number is large. A variety of numerical methods have been developed to obtain better approximations; and many of these methods fall under the generic classification of Eulerian-Lagrangian methods [32, 34, 63, 29, 8]. Among them, a family of locally conservative Eulerian-Lagrangian methods (LCELM) was introduced [31] in the simulation of immiscible displacement in porous media; extensive computational experiments were presented in [31, 30, 3, 2]. Optimal-order error estimates have been derived for a finite difference analogue of LCELM for a semilinear parabolic equation in a single space variable [26] and for a multidimensional LCELM based on mixed finite elements for a semilinear parabolic equation [27].

Details on FD and LCELM discretization are provided in subsequent sections.

3.1. Finite difference approximations (FD). Consider first Richards' equation (18) and apply the FD method, fully implicit in time. Here for brevity we present the formulation for $d = 1$.

Discretize the spatial domain Ω as covered by a rectangular uniform grid of size $\triangle x$, with cell centers denoted by $\mathbf{x}_i$, and unknowns denoted by subscript $_i$, $i = 1, \dots nx$. The time variable is discretized by splitting $(0, T)$ into subintervals of variable size $\triangle T_n, n = 1, \dots N$ where N is the number of time steps. Here the values of unknowns at time t^n are denoted by superscripts ^{n}.

We set

$$S_i^0 := S_0(\mathbf{x}_i), \quad i = 1, \dots nx. \tag{26}$$

and then, for $n = 1, \dots N$, we solve

$$
\begin{aligned}
(27)\quad (\triangle x)^2\phi(S_i^n - S_i^{n-1}) & \\
& - \triangle T_n K_{i+1/2}\frac{1}{\mu}k^*_{w,i+1/2}\left(P^n_{i+1} - P^n_i\right) \\
& + \triangle T_n K_{i+1/2}\frac{1}{\mu}k^{**}_{w,i+1/2}\rho G\left(D_{i+1} - D_i\right) \\
& + \triangle T_n K_{i-1/2}\frac{1}{\mu}k^*_{w,i-1/2}\left(P^n_i - P^n_{i-1}\right) \\
& - \triangle T_n K_{i-1/2}\frac{1}{\mu}k^{**}_{w,i-1/2}\rho G\left(D_i - D_{i-1}\right) = 0,\ i = 1, \dots nx.
\end{aligned}
$$

In this scheme the main gist and the source of variants in formulations is in a) the choice of time-lagged coefficients or fully implicit solution, b) the handling of the advective $k^{**}_{w,i\pm1/2}$ and the diffusive $k^*_{w,i\pm1/2}$ edge nonlinearities, and finally c) the choice of primary unknowns: P or S. The various variants are summarized in Table 1 with details given as follows. Our methods resemble most closely those in [40, 69, 45, 53]. Boundary conditions are treated as in [53].

These variants follow standard textbook procedures [52, 43] but their choices may be delicate especially in the forthcoming context of dynamic capillary pressure terms; we give details for completeness.

For a) time discretization we use time-lagging whereby we set $n* := n - 1$ or a fully implicit solution in which we use $n* := n$. These definitions are used in computing coefficients of the (non)linear system.

In b) the choices include upwinding denoted in Table 1 by 'u' or arithmetic averages denoted by 'a'. Specifically, in upwinding, we set $k^*_{w,i+1/2} = k_w(S_i^{n*})$ provided the potential difference $\triangle\psi^{n*}_{i+1/2} := P^{n*}_{i+1} - P^{n*}_i - \rho G(D_{i+1} - D_i)$ is negative indicating that the flow is from the left. If the potential difference indicates flow from the right, we use the value $k^*_{w,i+1/2} = k_w(S^{n*}_{i+1})$. In arithmetic averaging, we select $k^*_{w,i+1/2} = \frac{1}{2}(k_w(S_i^{n*}) + k_w(S^{n*}_{i+1}))$. Handling of $k^{**}_{w,i+1/2}$ and all other terms is analogous. The upwinding, albeit associated with lower convergence rate, provides additional stability and is applicable to more general models with compressibility and multiple rock types; see discussion in [40].

As concerns c), in the model (22) preferred by hydrologists and solved for S only, the term $\mathbf{D}(S)$ given by (20) is discretized using arithmetic averaging on the *entire* term and the term $\mathbf{C}(S)$ is discretized using upwinding as in all other FD variants. It appears that the averaging of the lumped form of $\mathbf{D}(S)$ leads to faster convergence than without the chain rule.

In the original model (18), when it is solved for P, the values of S and consequently the nonlinear properties are obtained via (19). Depending on the choice of time-lagging or fully implicit solution, the relation (19) is applied at the same time step or with time lagging.

Each of these choices has associated existing numerical theory which applies as long as the values of $k_w(\mathbf{x}, S)$, $P_c'(S)$ (or $\mathbf{D}(S)$) remain bounded away from singularities. In particular, it is known that for large advection, the use of arithmetic averaging in $k^{**}_{w,i+1/2}$ leads to instabilities. On the other hand, the use of upwinding increases numerical diffusion and leads to schemes that are slightly less accurate, at least away from degenerate conditions [48].

3.1.1. Discretization of two-phase model (3)–(4)**.** In analogy to what was done for Richards' equation; that is, (3), one can write the discrete analogue for the nonwetting phase (4) and consider multiple variants of discretizations. For simplicity, here we report only on those discretization variants in which we use upwinding for both convective and diffusive terms. Time discretization can be implicit or time-lagged, the unknowns chosen are P, S.

3.1.2. Incorporating dynamic capillary pressure. In order to model dynamic capillary pressure, we discretize (23) as follows:

$$P_n^{n*} - P^{n*} = P_c(\mathbf{x}, S^{n*}) - \tau \frac{S^n - S^{n-1}}{\triangle T_n}, \tag{28}$$

where for implicit solution we use $n* = n$ and for time-lagging we use $n* = n-1$.

In the case of Richards' equation we use $P_n^{n*} \equiv 0$ from which follows a modification of (28).

3.1.3. Solution of (27)**.** The system of discrete equations (27) for the time-lagged case when $n* \equiv n-1$ requires application of a linear solver. We use a direct solver from MATLAB suite for sparse matrices or an SOR iteration in which we iterate to very small tolerance.

In the fully implicit case when $n* \equiv n$ we have a nonlinear system to solve at every time step. This is done by Newton's iteration with the Jacobian computed analytically and with the initial guess extrapolated from previous time steps. The time step is either fixed or it is controlled automatically depending on the success or failure of the Newton iteration. We refer to [46] for a general reference to solving nonlinear systems with Newton's method and to [44, 25, 45] for the work applicable to unsaturated flow.

3.2. Locally conservative Eulerian-Lagrangian method. LCELM is based on an operator-splitting procedure that separates the transport (convection) from the diffusion in (24) as follows:

1° **Initialize:**

$$S^0(\mathbf{x}) = S(\mathbf{x}, 0). \tag{29}$$

2° **Transport(Gravity):** For $n \geq 1$,

$$\frac{\partial(\phi \overline{S})}{\partial t} + \nabla \cdot \left(\frac{\rho_w G}{\mu} \mathbf{K} k_w(\overline{S}) \nabla D \right) = 0, \quad t^{n-1} < t < t^n, \tag{30a}$$

$$\overline{S}(\mathbf{x}, t^{n-1}) = S^{n-1}(\mathbf{x}). \tag{30b}$$

3° **Diffusion(Capillary pressure):** For $n \geq 1$,

$$\frac{\partial(\phi \widehat{S})}{\partial t} + \nabla \cdot \left(\frac{1}{\mu} \mathbf{K} k_w(\widehat{S}) \left(P_c'(\widehat{S}) \nabla \widehat{S} - \nabla \tau \frac{\partial \widehat{S}}{\partial t} \right) \right) = 0, \quad t^{n-1} < t < t^n, \tag{31a}$$

$$\widehat{S}(\mathbf{x}, t^{n-1}) = \overline{S}(\mathbf{x}, t^n). \tag{31b}$$

4° Set

$$S^n(\mathbf{x}) = \widehat{S}(\mathbf{x}, t^n), \tag{32}$$

and go to 2°.

3.2.1. The local conservation relation. Let $M_{ij}, i = 1, \ldots nx, j = 1, \ldots ny$ be the $\triangle x \times \triangle y$ rectangle given by

$$M_{ij} = [\mathbf{x}_{i-1/2,j-1/2},\, \mathbf{x}_{i+1/2,j-1/2}] \times [\mathbf{x}_{i-1/2,j-1/2}, \mathbf{x}_{i-1/2,j+1/2}].$$

Now, let us define a predecessor set for M_{ij}. Let $\mathbf{y}(t;\mathbf{x})$ be the solution of the final value problem given by

$$\mathbf{y}' = \frac{\rho_w G \mathbf{K} k_w(S) \nabla D}{\mu \phi S}, \tag{33a}$$

$$\mathbf{y}(t^n, \mathbf{x}) = \mathbf{x}. \tag{33b}$$

Then, let

$$\partial \tilde{M}_{ij}^n = \{\mathbf{y}(t^{n-1};\mathbf{x}) : \mathbf{x} \in \partial M_{ij}\}, \tag{34}$$

and define the interior of $\partial \tilde{M}_{ij}^n$ to be the predecessor set $\tilde{M}_{ij}$ at time t^{n-1} corresponding to M_{ij} at time t^n. Define the tube E_{ij}^n to be the set interior to $M_{ij}, \tilde{M}_{ij}^n$ and the lateral boundary F_{ij}^n defined by the integral curve $\mathbf{y}(t;\mathbf{x}), t^{n-1} < t < t^n$, $\mathbf{x} \in \partial M_{ij}$. Then, the solution of (24) satisfies the relation

$$\int_{M_{ij}} \phi S^n \, d\mathbf{x} - \int_{\tilde{M}_{ij}} \phi S^{n-1} \, d\mathbf{x} + \int_{E_{ij}} \nabla \cdot \left(\frac{1}{\mu} \mathbf{K} k_w(\mathbf{x}, S) \left(P_c'(S) \nabla S - \nabla \tau \frac{\partial S}{\partial t} \right) \right) d\mathbf{x} dt = 0. \tag{35}$$

3.2.2. The approximate transport. The sets $\tilde{M}_{ij}^n$ and E_{ij}^n depend explicitly on the solution S of (22), thus, they must be approximated using the values of the approximate solution, which will be denoted by S^n, $n = 0, 1 \ldots N$. In this paper, we shall limit ourselves to the lowest order version of the LCELM based on a finite difference method in that S^n is taken to be piecewise constant. Note that S^n is multivalued on ∂M_{ij}, consequently, we will introduce a piecewise bilinear interpolation operator I as follows:

$$IS(\mathbf{x}_{i-1/2,j-1/2}) = (S_{i-1,j-1} + S_{i,j-1})/2, \tag{36}$$

i.e., IS is the upstream average. Now, assume that S^{n-1} is known. Denote the vertices of M_{ij} by $\mathbf{x}_{ij,k}$, $k = 1, \ldots, 4$. Then, define an approximate predecessor $\widehat{Q}_{ij}^n$ as the quadrilateral having vertices

$$\widehat{\mathbf{x}}_{ij,k}^n = \mathbf{x}_{ij,k} - \frac{\rho_w G \mathbf{K} k_w(IS(\mathbf{x}_{ij,k})) \nabla D(\mathbf{x}_{ij,k})}{\mu \phi IS(\mathbf{x}_{ij,k})} \triangle T, \quad k = 1, \ldots, 4. \tag{37}$$

Let $\widehat{E}_{ij}^n$ be the tube formed with top M_{ij} and bottom $\widehat{Q}_{ij}^n$. Then, the approximate local conservation equation can be written as

$$\int_{M_{ij}} \overline{S}^n \, d\mathbf{x} = \int_{\widehat{Q}_{ij}^n} S^{n-1} \, d\mathbf{x}. \tag{38}$$

3.2.3. The diffusive fractional step. We shall approximate the solution of the diffusive fractional step (31) by one of two variants of time-lagged cell-centered finite difference methods, referred to below by R_LCELM and R_LCELM2. Here,

we for simplicity assume that τ is a constant and define:

$$
\begin{aligned}
(39)\qquad & (\triangle x)^2(\triangle y)^2\phi(S^n_{i,j}-\overline{S}^n_{i,j}) \\
&+\frac{(\triangle y)^2}{\mu}\mathbf{K}_{i+1/2,j}\Big\{\triangle T(k_wP_c')^{*,n}_{i+1/2,j}(S^n_{i+1,j}-S^n_{i,j}) \\
&\quad -\tau k^{*,n}_{w,i+1/2,j}\Big((S^n_{i+1,j}-S^n_{i,j})-(\tilde{S}^n_{i+1,j}-\tilde{S}^n_{i,j})\Big)\Big\} \\
&-\frac{(\triangle y)^2}{\mu}\mathbf{K}_{i-1/2,j}\Big\{\triangle T(k_wP_c')^{*,n}_{i-1/2,j}(S^n_{i,j}-S^n_{i-1,j}) \\
&\quad -\tau k^{*,n}_{w,i-1/2,j}\Big((S^n_{i,j}-S^n_{i-1,j})-(\tilde{S}^n_{i,j}-\tilde{S}^n_{i-1,j})\Big)\Big\} \\
&+\frac{(\triangle x)^2}{\mu}\mathbf{K}_{i,j+1/2}\Big\{\triangle T(k_wP_c')^{*,n}_{i,j+1/2}(S^n_{i,j+1}-S^n_{i,j}) \\
&\quad -\tau k^{*,n}_{w,i,j+1/2}\Big((S^n_{i,j+1}-S^n_{i,j})-(\tilde{S}^n_{i,j+1}-\tilde{S}^n_{i,j})\Big)\Big\} \\
&-\frac{(\triangle x)^2}{\mu}\mathbf{K}_{i,j-1/2}\Big\{\triangle T(k_wP_c')^{*,n}_{i,j-1/2}(S^n_{i,j}-S^n_{i,j-1}) \\
&\quad -\tau k^{*,n}_{w,i,j-1/2}\Big((S^n_{i,j}-S^n_{i,j-1})-(\tilde{S}^n_{i,j}-\tilde{S}^n_{i,j-1})\Big)\Big\} \\
&\qquad =0,
\end{aligned}
$$

where $(k_wP_c')^{*,n}_{i\pm1/2,j}$ and $k^{*,n}_{w,i\pm1/2,j}$ are the arithmetic mean,

$$(k_wP_c')^{*,n}_{i\pm1/2,j}=\Big((k_wP_c')(\overline{S}^n_{i,j})+(k_wP_c')(\overline{S}^n_{i\pm1,j})\Big)/2$$

and

$$k^{*,n}_{w,i\pm1/2,j}=\Big(k_w(\overline{S}^n_{i,j})+k_w(\overline{S}^n_{i\pm1,j}\Big)/2,$$

respectively and $(k_wP_c')^{*,n}_{i,j\pm1/2}$ and $k^{*,n}_{w,i,j\pm1/2}$ are defined analogously. Here,

$$\tilde{S}^n_{i,j}=\begin{cases}\overline{S}^n_{i,j} & \text{in R\_LCELM,}\\ S^{n-1}_{i,j} & \text{in R\_LCELM2.}\end{cases}\qquad(40)$$

A convergence analysis for R_LCELM2 for a linear problem has been done and will be presented in a subsequent paper.

4. Results

Here we report on numerical results for the schemes presented above. First, we illustrate the convergence and stability discussion given above; this is done via Example 1 in which no dynamic capillary pressure is taken into account i.e. $\tau\equiv 0$. Next, in Example 2 we compare the solutions with and without dynamic capillary pressure and are concerned with convergence and stability of the schemes. In Example 3 we demonstrate the use of multiple rock types combined with dynamic capillary pressure using a two-phase formulation.

Examples 1 and 2 are essentially 1D even though they are run with 2D codes; example 3 is run essentially in 3D but it has 2D features only.

Our studies are comprehensive and include simulations across all listed cases and schemes. However, for brevity we present only the typical and/or most interesting cases.

4.1. Example 1: convergence studies for $\tau \equiv 0$. Our examples include i) horizontal (1) and ii) vertical (2) infiltration problem, with parameters shown in Table 4.1. Initial saturation S_0 is constant, infiltration proceeds from the left at S_D. This is a purely "static" case with $\tau = 0$. Its purpose is to demonstrate convergence of methods which are used in Example 2 to simulate dynamic capillary pressure phenomena.

The data S_0, S_D for two-phase flow model are chosen so that (17) is essentially satisfied. Note, however, that this won't happen for just any choice of values S_0, S_D . The data P_0, P_D are adjusted accordingly.

Overall, the simulation parameters in this Example are chosen to demonstrate robustness of the methods for the relatively hard case which is the infiltration of a wetting front into an initially very dry soil. The infiltration is due to a boundary condition on i) left and ii) top, respectively, in each case imposing a pressure gradient. The right (bottom) boundary conditions are the same as the initial condition.

Recall that i) means the problem has only nonlinear (degenerate) diffusion terms while in case ii) it has additional nonlinear convective terms. Consulting Table 4.1 and Figure 1 we see that the case of sand II (fine) with large capillary pressure is diffusive, with or without gravity terms, while the case of sand I (coarse) has a strong convective term when gravity is present. This is evident in the solutions shown in Figure 4 which shows convergence study for Richards' equation and two-phase formulations.

It is clear that all methods converge, albeit R_FD1 and R_LCELM appear to have a higher convergence rate. Due to the upwinding R_FD2–R_FD5 have lower convergence rates. Moreover, all schemes and both models, Richards and two-phase, agree very well qualitatively and in most cases quantitatively. It seems that the choice of time-lagging or implicit solution is not essential as is the choice of the time step, as long as the time step is refined along with spatial grid parameter (studies not shown). Finally, what is not visible on the picture due to its resolution, is the fact that both methods R_LCELM and R_FD1 as opposed to R_FD2–R_FD5 appear to converge to a solution from *opposite* sides. This fundamental difference must be caused by the use of arithmetic averaging to *all* of $\mathbf{D}(S)$ in the former methods as opposed to handling them directly via a dependent variable in the latter case. In fact, the R_FD1 method appears to be closely related to standard Galerkin finite element discretization via the lumped discretization/averaging whereas the R_FD2–R_FD5 relate closer to the mixed methods.

Next, the choice of primary unknown: P or S in Richards' formulation with implicit time stepping appears to suggest, perhaps not surprisingly, that the use of P (implicit S) allows for larger time steps and smoother convergence of Newton's iteration than the use of S which is less diffusive.

More examples and comparisons are given in Figure 5 where we show results for multiple numerical methods for a given discretization, with gravity. Here we see again that R_LCELM and R_FD1 are capable of producing sharper fronts than R_FD2–R_FD5. The case without gravity is not shown but it features all curves in essentially the same place for all schemes.

Finally, as concerns comparison between modeling unsaturated flow using Richards' or two-phase flow model, it appears that for those cases when $P_n \approx 0$, the differences between (almost) equivalent methods R_FD2–R_FD5 and 2PH* are negligible. Here and below, 2PH* represents all the variants of FD based on the two-phase model.

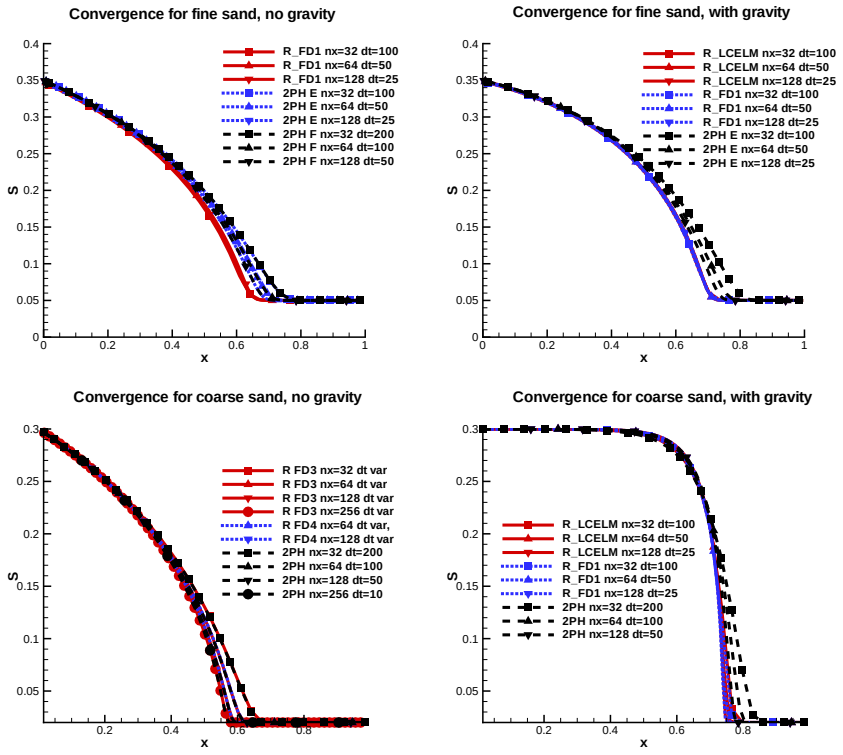


FIGURE 4. Convergence studies for fine sand II (top) and coarse sand I (bottom) examples, for different numerical methods; shown are cases with i) no convection (left, horizontal flow (1)) and ii) with convection (right, vertical flow (2)). Here nx varies from 32 to 256, and $\triangle T$ is changed appropriately.

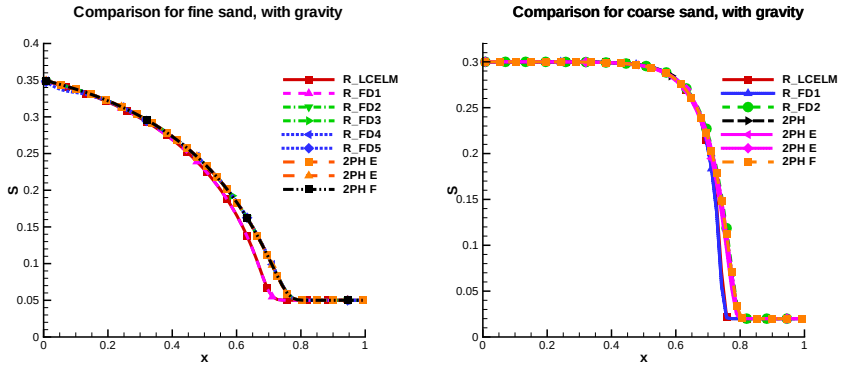


FIGURE 5. Comparison between numerical methods used for Richards' equation and two-phase flow formulation for both types of sand, with gravity.

4.2. Example 2: implementation of dynamic capillary pressure. Here we continue simulations with scenarios as in Example 1. However, now the dynamic

Physical properties of fluids:
$\rho_w = 10^3[kg/m^3]$, $\rho_n = 1.2[kg/m^3]$,
$\mu_w = 10^{-3}[Pas]$, $\mu_n = 1.78 \cdot 10^{-5}[Pas]$,
$G = 9.8066[m/s^2]$.

Properties of porous medium:
$\phi = 0.4$,
$K^I = 10^{-10}[m^2]$,
$K^{II} = 10^{-11}[m^2]$.

van Genuchten parameters:
coarse sand I: n = 2.494, $\alpha = 10^{-3}$,
fine sand II: n = 2.237, $\alpha = 10^{-4}$,
all cases: $\epsilon = 0.5$, $\gamma = 1/3$.

Boundary and initial conditions:
coarse sand I: $S_0 = 0.02, S_{left} = 0.30$,
fine sand II: $S_0 = 0.05, S_{left} = 0.35$.

TABLE 2. Simulation parameters for Example 1 and Example 2, in SI units. The porous medium and rock-fluid properties are similar to those in the literature [43, 42]. We focus on effective saturations, therefore $S_{wr} = 0$.

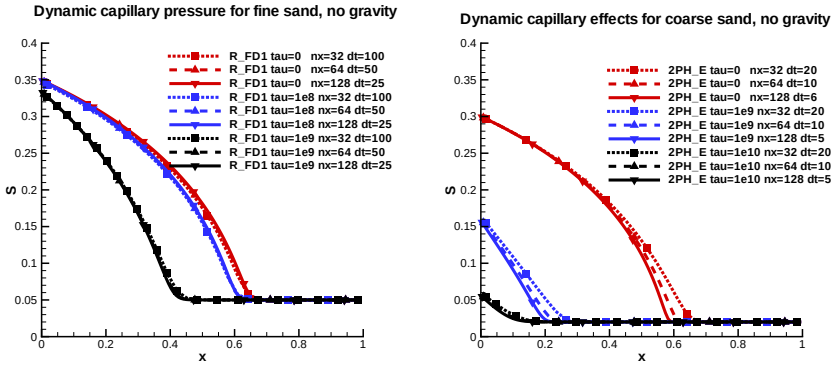


FIGURE 6. Influence of dynamic capillary pressure and convergence tests, no gravity; both fine and coarse sand cases.

capillary pressure terms are present, that is, $\tau > 0$. To a numerical analyst it actually comes as a surprise that in order to see the influence of dynamic capillary pressure, the values of τ (in the units used) have to be of several orders of magnitude, see however the identification of τ in [42].

Figures 6–9 present results of simulation for the case from Example 1, except with $\tau > 0$.

The first observation is that all methods (some cases are not shown) appear to converge and have smooth solutions; at least as long as no significant convection is present. We see again higher convergence rates for R_LCELM and R_FD1 and lower

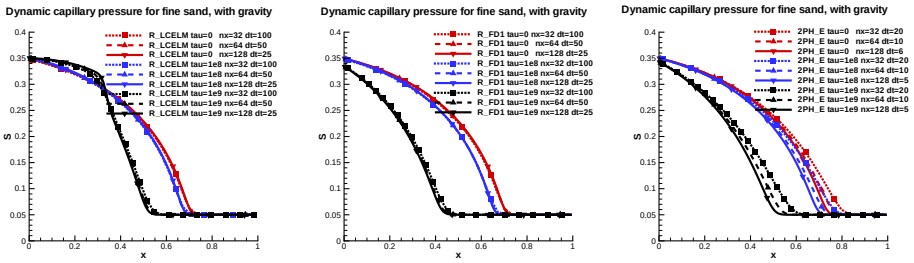


FIGURE 7. Convergence studies with dynamic capillary pressure: Fine sand with gravity (some convection, but not dominating). Solution profiles for R_LCELM2 are essentially identical with those for R_FD1. Relative error, $\|R\_FD1 - R\_LCELM2\|_{\ell^\infty}/\|R\_FD1\|_{\ell^\infty}$ is less than 0.16% for each τ when $nx = 128$, $dt = 25$.

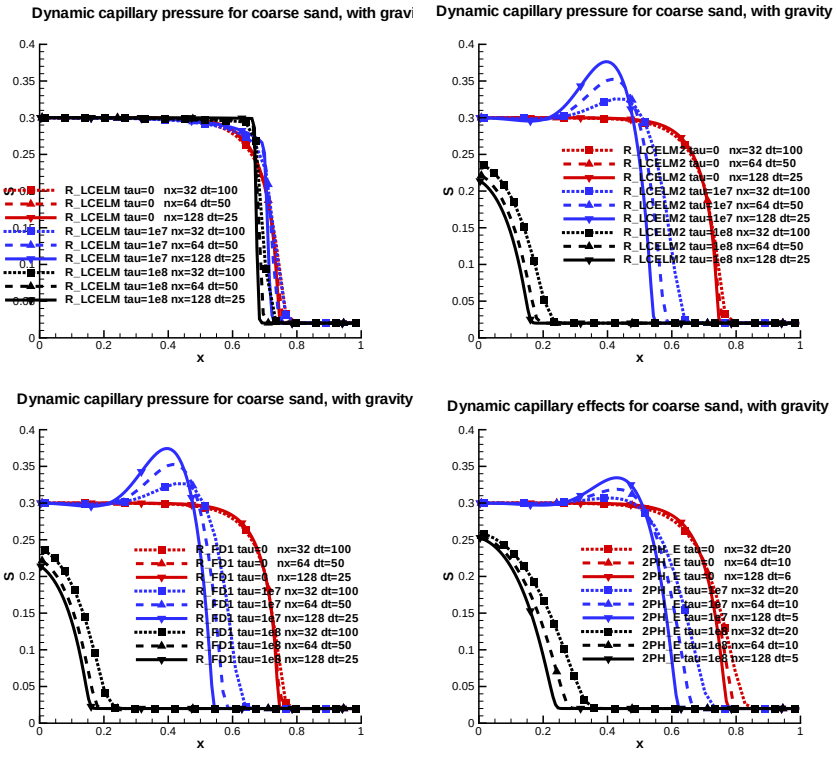


FIGURE 8. Convergence studies with dynamic capillary pressure. Coarse sand with gravity (very significant convection). Relative error, $\|R\_FD1 - R\_LCELM2\|_{\ell^\infty}/\|R\_FD1\|_{\ell^\infty}$ is less than 6.8% for each τ when $nx = 128$, $dt = 25$.

for R_LCELM2, R_FD2–R_FD5 and 2PH*. This is consistent with the regularity theory presented in [59] which suggests that the presence of dynamic capillary pressure terms should not have de-stabilizing effects, at least in the purely parabolic case where convection is absent.

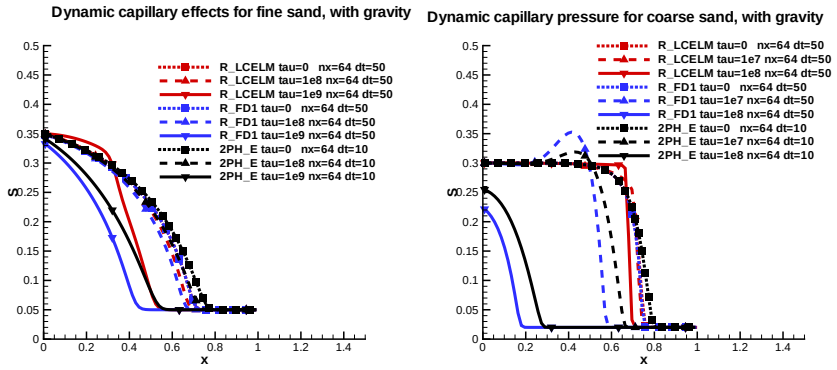


FIGURE 9. Comparison of results for dynamic capillary pressure. Both sand types with gravity, various methods.

Now, with some convection present (Figure 7), the convergence with the same observations is apparent. However, R_LCELM leads to a sharpening of the front, while R_LCELM2 doesn't. This can be explained by ability of LCELM to handle transport regardless of direction. The sharpening of the front visible in the R_LCELM case can be explained by the effect of terms with large τ which lead to a decrease of diffusion stronger for large S for this variant of time-stepping. See (40).

Then, for dominant convection (Figure 8), we obtain results which have nonmonotone profiles for large τ for *all* FD-based methods, regardless of whether the Richards' or two-phase flow formulation is used. Note also that R_LCELM2 presents qualitatively the same profiles as FD-based methods. It is hard to assess convergence in this case as we do not have the true solution at this point. Also, we cannot speculate whether the apparent nonmonotonicity of profiles in Figure 8, present for all FD and R_LCELM2 solutions, relates to a numerical instability, or to a physical phenomenon. One could speculate that R_LCELM profiles are "real" and therefore instabilities in FD arise due to their lack of ability to approximate a sharp front which arises "against" the apparent convection direction. One could also bring back the discussion of nonmonotonicity in [40] for a vertical infiltration problem with heterogeneity.

To put these results in perspective, we recall that it has been reported in [42] that a straightforward discretization of the dynamic capillary pressure term leads to instabilities for large values of τ, albeit the details of the problematic and of improved formulation(s) were not given. On the other hand, the results reported in [51] focus on instabilities in the similarity solutions due to dynamic capillary pressure terms varying in t which are argued to be physical and not merely numerical phenomena.

Our results confirm that the straightforward inclusion of the dynamic capillary pressure does not lead to instabilities when convection terms are not present or are not dominant. At the same time, it is clear that the presence of advective terms which dominate diffusion may lead to physical appearance of various internal and boundary layers which may perhaps explain the nonmonotonicity of the solutions reported in Figure 8. At this point however we are not ready to make more general conclusions without further analysis which is outside the scope of this paper.

We close this discussion with remarks on time stepping and choice of primary unknowns. For all FD cases, the presence of dynamic terms appears to slow down the dynamics of the flow, similarly as reported in [42]. From a numerical point of view, this results in smoother performance of implicit methods, and very large time steps can be taken without harming the convergence of the Newton iteration. For example, the automatic time-stepping, if unrestricted, would allow the time step to grow by three orders of magnitude. However, the use of S as primary unknown in implicit formulations of Richards' equation required very small time steps as the Newton iteration was very sensitive to the domain of validity of (23).

The qualitative behavior of solutions for increasing τ comes without surprise. As shown in Figure 2, the larger dynamic terms during imbibition decrease the apparent diffusion due to capillary pressure, hence they should slow down the front which is moving partly due to nonlinear diffusion, and partly to convection. Where the two effects (diffusion and convection) become comparable, it is perhaps their competition that leads to instabilities which cannot be reliably captured with the numerical methods discussed here.

4.3. Example 3: dynamic capillary pressure and different rock types. In this experiment we show results of simulation for a 20×20 cell with $D(\mathbf{x}) = 0$ of heterogeneous medium such as shown in Figure 3. We are interested in the combined effects of heterogeneity and dynamic capillary pressure. The numerical scheme we choose is the implicit implementation of two-phase model with upwinding and variable time-stepping listed as 2PH in Table 1.

The data used for this example is very simple and except for the special data described below it is as in Table 4.1. Heterogeneity is associated with the ratio of 3 orders of magnitude difference in permeabilities $\mathbf{K}^I = 10^3 \cdot \mathbf{K}^{II}$. We use $k_w(\mathbf{x}, S) = S^2, k_n(\mathbf{x}, S) = (1-S)^2$ for both rock types. As concerns static and dynamic capillary pressure relationships, we consider four experiments. The first two are with static capillary pressure and read a) $P_c^I(S) \equiv P_c^{II}(S) = \frac{1}{10\sqrt{S}}$; $\tau_I = 0$; $\tau_{II} = 0$, b) $P_c^I(S) = \frac{1}{10\sqrt{S}}$; $P_c^{II}(S) = 0.5(1-S)$; $\tau^I = 0$; $\tau^{II} = 0$. In next two experiments we vary dynamic capillary pressure coefficients c) $P_c^I(S) = \frac{1}{10\sqrt{S}}$; $P_c^{II}(S) = 0.5(1-S)$; $\tau^I = 10$; $\tau^{II} = 0$, d) $P_c^I(S) = \frac{1}{10\sqrt{S}}$; $P_c^{II}(S) = 0.5(1-S)$; $\tau^I = 0$; $\tau^{II} = 10$. That is, in general, we have

$$P_c^I(S) \neq P_c^{II}(S). \tag{41}$$

All examples start from an initial equilibrium in which we are given a constant equilibrium pressure in both phases (hence, no pressure gradient) across both rock types. Such an equilibrium implies equality of capillary pressures and, in the case of (41), this implies inequality of initial saturations; see the initial condition of S for cases b), c), and d) shown below. On right and left boundaries of the cell we apply a Dirichlet boundary condition for saturations and pressures and no-flow condition is used on remaining boundaries. Thereby we create an infiltration front moving from right to left and which results in appearance of some internal boundary layers close to the outlet boundary.

Results of simulations are shown in Figures 10 and 11. They show the importance of both heterogeneity and dynamic effects.

In particular, comparison in Figure 10 shows significance of (41) starting at initial time step and in what follows. Case a) shows an example of bypassed air (nonwetting phase) pockets inside of the cell where the wetting phase has not invaded.

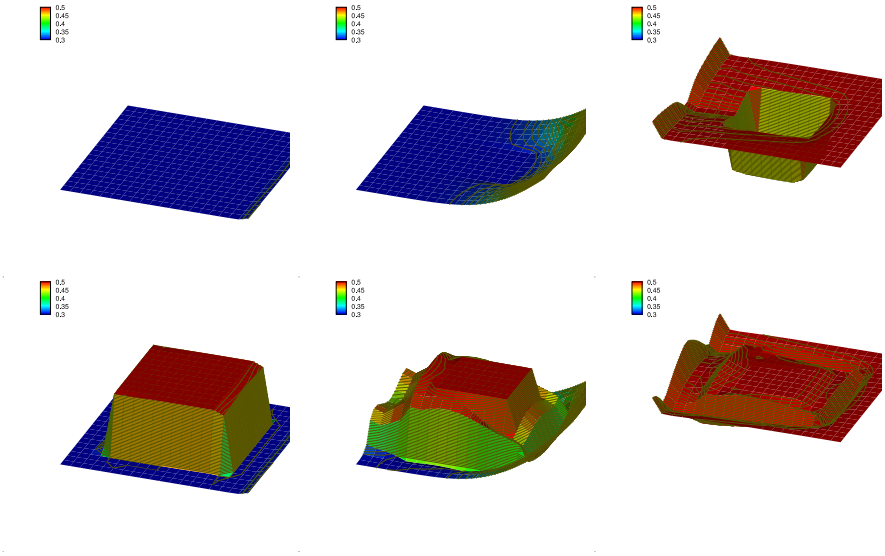


FIGURE 10. Results with static capillary pressure for cases a) (top) and b) (bottom), at the beginning, middle, and last time step of the simulation (from left to right).

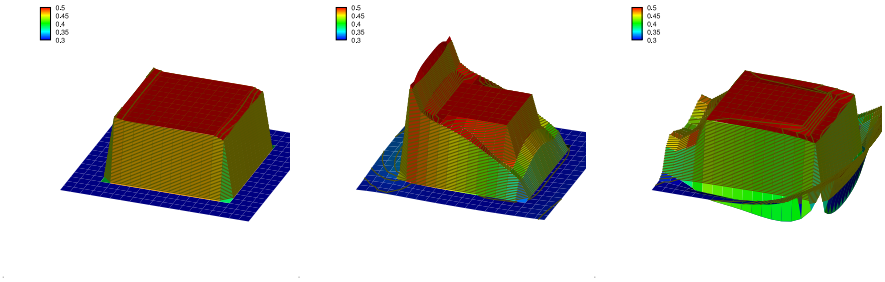


FIGURE 11. Initial time step (left) for cases c) and d) and results with dynamic capillary pressure at the end of simulation: case c) (middle) and d) (right).

Comparison of cases b) with dynamic cases of c) and (more physical) d) reveals that the delay effects in sand I (case c) slow down the flow in Ω^I surrounding Ω^{II} and lead to an internal boundary layer close to left boundary of Ω^{II}. The effect is reversed in case d). We mention here that it appears (Majid Hassanizadeh, private communications) that case d) is more physical: dynamic effects are likely to be more significant in fine sand where capillary pressure is more important than in coarse where they are not as important.

Characteristics of case d) will be the building blocks in our future construction of multiscale models of preferential flow with dynamic effects.

5. Conclusions

While the rigorous analysis for the nonlinear degenerate case is outside the scope of this paper, the convergence of FD formulation for (22) with dynamic capillary pressure (pseudo-parabolic) terms without convection i.e. $\mathbf{C}(S) \equiv 0$ has been confirmed by numerical experiments and is consistent with the findings of [35].

As reported above for case i) without gravity all the FD-based methods appear to converge and to have monotone solutions. For this case however the LCELM formulation is not relevant.

In the general case, we have shown above that in the presence of dynamic capillary pressure as well as different rock types, numerical solutions exhibit substantial differences with respect to those without dynamic terms. In cases without convection such as (22) with $\mathbf{C}(S) \equiv 0$, the solutions for large τ lag behind. This is consistent with the analysis in [60, 58] predicting that the size of jump in initial data, here due to the difference between initial and boundary conditions, decreases slowly when large τ in pseudo-parabolic terms is used.

However, it appears that the main difficulties and the presence of apparent instabilities are associated with the strong convective terms $\mathbf{C}(S)$ and large τ. In separate experiments not reported here we were able to determine, for each τ, the critical size of D_{fac} in $D(\mathbf{x}) = D_{fac}\mathbf{x}$ for which the method remains stable. In other words, there appears a critical Péclet number beyond which, due to the boundary or internal layers the numerical solution exhibits nonmonotonicities which remain stable with respect to the spatial and temporal grid refinement.

As concerns the convective term, our hope was that the use of LCELM would alleviate any potential instabilities. Indeed, the R_LCELM (but not R_LCELM2) results appear stable and have monotone behavior which is stable with respect to mesh refinement. While this approach offers different results than FD-based methods, we believe that more analysis of the time splitting and of the influence of convective terms is necessary before firm conclusions are drawn as to the nature and convergence of the methods.

Next, we believe that one needs a two-phase model rather than Richards' equation to properly model both the heterogeneities and dynamic capillary pressure effects. This follows from our observations on the size of nonwetting phase pressure P_n as monitored in the two-phase flow model which, albeit small compared to the value of P, exhibits substantial variations especially in dynamic case.

Finally, as seen from Example 3, the numerical methods for preferential flow should take into account both the variation rock type as well as proper models of accounting for dynamic effects such as dynamic capillary pressure. The impact of these two elements on the solutions is substantial, both qualitatively and quantitatively.

Acknowledgements

We would like to thank Majid Hassanizadeh for giving us the references to work [49] which pointed us also to other works in that volume, and for fruitful discussions. Most of our research was done before we became aware of these additional references. We thank Ralph Showalter and John Selker for discussions and references, and inspiration.

The research of the first author was partially supported by the grants NSF grant 0511190 and DOE grant 98089; the second author was partially supported by DOE grant 98089.

References

[1] M. Afif and B. Amaziane. Convergence of finite volume schemes for a degenerate convection-diffusion equation arising in flow in porous media. *Comput. Methods Appl. Mech. Engrg.*, 191(46):5265–5286, 2002.

[2] César Almeida, Jim Douglas, Jr., and Felipe Pereira. A new characteristics-based numerical method for miscible displacement in heterogeneous formations. *Comput. Appl. Math.*, 21(2):573–605, 2002. Special issue on multi-scale science (Nova Friburgo, 2000).

[3] César Almeida, Jim Douglas, Jr., Felipe Pereira, Luis Carlos Roman, and Li-Ming Yeh. Algorithmic aspects of a locally conservative Eulerian-Lagrangian method for transport-dominated diffusive systems. In *Fluid flow and transport in porous media: mathematical and numerical treatment (South Hadley, MA, 2001)*, volume 295 of *Contemp. Math.*, pages 37–48. Amer. Math. Soc., Providence, RI, 2002.

[4] H. W. Alt and E. di Benedetto. Nonsteady flow of water and oil through inhomogeneous porous media. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 12(3):335–392, 1985.

[5] H.W. Alt, S. Luckhaus, and A. Visintin. On nonstationary flow through porous media. *Ann. Mat. Pura Appl.*, 136(4):303–316, 1984.

[6] T. Arbogast. The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow. *Nonlinear Analysis, Theory, Methods and Applications*, 19:1009–1031, 1992.

[7] T. Arbogast, A. Chilakapati, and M. F. Wheeler. A characteristic-mixed method for contaminant transport and miscible displacement. In *Computational methods in water resources, IX, Vol. 1 (Denver, CO, 1992)*, pages 77–84. Comput. Mech., Southampton, 1992.

[8] T. Arbogast and M. F. Wheeler. A characteristics-mixed finite element method for advection-dominated transport problems. *SIAM J. Numer. Anal.*, 32:404–424, 1995.

[9] Todd Arbogast, Mary F. Wheeler, and Nai-Ying Zhang. A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.*, 33(4):1669–1687, 1996.

[10] Douglas N. Arnold, Jim Douglas, Jr., and Vidar Thomée. Superconvergence of a finite element approximation to the solution of a Sobolev equation in a single space variable. *Math. Comp.*, 36(153):53–63, 1981.

[11] G. I. Barenblatt, D. B. Silin, and T. W. Patzek. The mathematical model of non-equilibrium effects in water-oil displacement. *SPEJ*, 8(4):409–416, December 2003.

[12] G.I. Barenblatt and A. A. Gil'man. A mathematical model of non-equilibrium counter-current capillary imbibition. *J. Eng. Phys.*, 52(3):456–461, 1987.

[13] A. Y. Beliaev and R. J. Schotting. Analysis of a new model for unsaturated flow in porous media including hysteresis and dynamic effects. *Comput. Geosci.*, 5(4):345–368 (2002), 2001.

[14] A. Yu. Beliaev and S. M. Hassanizadeh. A theoretical model of hysteresis and dynamic effects in the capillary relation for two-phase flow in porous media. *Transp. Porous Media*, 43(3):487–510, 2001.

[15] L. Bergamaschi and M. Putti. Mixed finite elements and Newton-type linearizations for the solution of Richards' equation. *Internat. J. Numer. Methods Engrg.*, 45(8):1025–1046, 1999.

[16] Alain Bourgeat. Two-phase flow. In U. Hornung, editor, *Homogenization and porous media*, volume 6 of *Interdiscip. Appl. Math.*, pages 97–187. Springer, New York, 1997.

[17] Alain Bourgeat and Abdelkader Hidani. Effective model of two-phase flow in a porous medium made of different rock types. *Appl. Anal.*, 58(1-2):1–29, 1995.

[18] Alain Bourgeat and Abdelkader Hidani. A result of existence for a model of two-phase flow in a porous medium made of different rock types. *Appl. Anal.*, 56(3-4):381–399, 1995.

[19] Alain Bourgeat and Mikhail Panfilov. Effective two-phase flow through highly heterogeneous porous media: capillary nonequilibrium effects. *Comput. Geosci.*, 2(3):191–215, 1998.

[20] M. A. Celia, E. T. Bouloutas, and R. L. Zarba. A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, 26:1483–1496, 1990.

[21] Z. Chen, R. E. Ewing, Q. Jiang, and A. M. Spagnuolo. Error analysis for characteristics-based methods for degenerate parabolic problems. *SIAM J. Numer. Anal.*, 40(4):1491?–1515, 2002.

[22] Zhangxin Chen. Degenerate two-phase incompressible flow. I. Existence, uniqueness and regularity of a weak solution. *J. Differential Equations*, 171(2):203–232, 2001.

[23] Zhangxin Chen. Degenerate two-phase incompressible flow. II. Regularity, stability and stabilization. *J. Differential Equations*, 186(2):345–376, 2002.

[24] R. E. Collins. *Flow of fluids through porous materials*. PennWell Books, Tulsa, Oklahoma, 1961.

[25] C. N. Dawson, H. Klie, M. F. Wheeler, and C. Woodward. A parallel, implicit, cell-centered method for two-phase flow with a preconditioned Newton-Krylov solver. *Comput. Geosci.*, 1:215–249, 1997.

[26] J. Douglas, Jr. and C.-S. Huang. The convergence of a locally conservative eulerian-lagrangian finite difference method for a semilinear parabolic equation. *BIT*, pages 980–989, 2001.

[27] J. Douglas, Jr., A. M. Spagnuolo, and S.-Y. Yi. The convergence of a multidimensional locally conservative, Eulerian-Lagrangian finite element method for a semilinear parabolic equation. manuscript, 2007.
[28] Jim Douglas, Jr. and T. Arbogast. Dual-porosity models for flow in naturally fractured reservoirs. In J. H. Cushman, editor, *Dynamics of Fluids in Hierarchical Porous Media*, pages 177–221. Academic Press, 1990.
[29] Jim Douglas, Jr., Chieh-Sen Huang, and Felipe Pereira. The modified method of characteristics with adjusted advection. *Numer. Math.*, 83(3):353–369, 1999.
[30] Jim Douglas, Jr., Felipe Pereira, and Li-Ming Yeh. A locally conservative Eulerian-Lagrangian method for flow in a porous medium of a mixture of two components having different densities. In *Numerical treatment of multiphase flows in porous media (Beijing, 1999)*, volume 552 of *Lecture Notes in Phys.*, pages 138–155. Springer, Berlin, 2000.
[31] Jim Douglas, Jr., Felipe Pereira, and Li-Ming Yeh. A locally conservative Eulerian-Lagrangian numerical method and its application to nonlinear transport in porous media. *Comput. Geosci.*, 4(1):1–40, 2000.
[32] J. Douglas, Jr. and T. F. Russell. Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. *SIAM J. Numer. Anal.*, 19:871–885, 1982.
[33] C.J. Van Duijn, J. Molenaar, and M.J. De Neef. The effect of capillary forces on immiscible two-phase flow in heterogeneous porous media. *Trans. Porous Media*, 21:71–93, 1995.
[34] R. E. Ewing, T. F. Russell, and M. F. Wheeler. Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics. *Comp. Meth. Appl. Mech. Eng.*, 47:73–92, 1984.
[35] Richard E. Ewing. Numerical solution of Sobolev partial differential equations. *SIAM J. Numer. Anal.*, 12:345–363, 1975.
[36] Robert Eymard, Michaël Gutnic, and Danielle Hilhorst. The finite volume method for Richards equation. *Comput. Geosci.*, 3(3-4):259–294 (2000), 1999.
[37] William H. Ford. Galerkin approximations to non-linear pseudo-parabolic partial differential equations. *Aequationes Math.*, 14(3):271–291, 1976.
[38] William H. Ford and T. W. Ting. Stability and convergence of difference approximations to pseudo-parabolic partial differential equations. *Math. Comp.*, 27:737–743, 1973.
[39] William H. Ford and T. W. Ting. Uniform error estimates for difference approximations to nonlinear pseudo-parabolic partial differential equations. *SIAM J. Numer. Anal.*, 11:155–169, 1974.
[40] P. A. Forsyth and M. C. Kropinski. Monotonicity considerations for saturated-unsaturated subsurface flow. *SIAM J. Sci. Comput.*, 18(5):1328–1354, 1997.
[41] W.G. Gray and S.M. Hassanizadeh. Paradoxes and realities in unsaturated flow theory. *Water Resour. Res.*, 27:1847–1854, 1991.
[42] S.M. Hassanizadeh, M.A. Celia, and H.K. Dahle. Dynamic effects in the capillary pressure-saturation relationship and their impacts on unsaturated flow. *Vadose Zone Journal*, 1:38–57, 2002.
[43] R. Helmig. *Multiphase flow and transport processes in the Subsurface*. Springer, 1997.
[44] E. W. Jenkins, C. T. Kelley, C. E. Kees, and C. T. Miller. An aggregation-based domain decomposition preconditioner for groundwater flow. *SIAM Journal on Scientific Computing*, 23(2):430–441, 2001.
[45] Jim E. Jones and Carol S. Woodward. Newton-krylov-multigrid solvers for large-scale, highly heterogeneous, variably saturated flow problems. *Advances in Water Resources*, 24:763–774, 2001.
[46] C. T. Kelley. *Iterative methods for linear and nonlinear equations*. SIAM, Philadelphia, 1995.
[47] R.J. Lenhard, M. Oostrom, and M.D. White. Modeling fluid flow and transport in variably saturated porous media with the STOMP simulator. 2. Verification and validation exercises. *Advances in Water Resources*, 18(6), 1995.
[48] Randall J. LeVeque. *Numerical methods for conservation laws*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1990.
[49] S. Manthey, S. M. Hassanizadeh, and R. Helmig. Macro-scale dynamic effects in homogeneous and heterogeneous porous media. *Transp. Porous Med.*, 58:121–145, 2005.
[50] Ghislain De Marsily. *Quantitative Hydrogeology: Groundwater Hydrology for Engineers*. Academic Press, 1986.
[51] J. L. Nieber, R. Z. Dautov, and A. G. Egorov. Dynamic capillary pressure mechanism for instability in gravity-driven flows: Review and extension to very dry conditions. In D. B.

Das and S. M. Hassanizadeh, editors, *Upscaling Multiphase Flow in Porous Media*, pages 147–172. Springer, 2005.

[52] D. W. Peaceman. *Fundamentals of numerical reservoir simulation*. Elsevier Scientfic Publishing Company, Amsterdam-Oxford-New York, first edition, 1977.

[53] M. Peszyńska, E. Jenkins, and M. F. Wheeler. Boundary conditions for fully implicit two-phase flow model. In Xiaobing Feng and Tim P. Schulze, editors, *Recent Advances in Numerical Methods for Partial Differential Equations and Applications*, volume 306 of *Contemporary Mathematics Series*, pages 85–106. American Mathematical Society, 2002.

[54] Florin Radu, Iuliu Sorin Pop, and Peter Knabner. Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation. *SIAM J. Numer. Anal.*, 42(4):1452–1478 (electronic), 2004.

[55] T. Russell. Formulation of a model accounting for capillary non-equilibrium in naturally fractured subsurface formations. In R. E. Ewing and D. Copeland, editors, *Proceedings of the Fourth Wyoming Enhanced Oil Recovery Symposium*, pages 103–114, 1988.

[56] T. F. Russell and M. F. Wheeler. Finite element and finite difference methods for continuous flows in porous media. In R. E. Ewing, editor, *The Mathematics of Reservoir Simulation*, pages 35–106. SIAM, Philadelphia, 1983.

[57] J.S. Selker, C. K. Keller, and J. T. McCord. *Vadose Zone Processes*. Lewsi Publishers, 1999.

[58] R. E. Showalter. Partial differential equations of Sobolev-Galpern type. *Pacific J. Math.*, 31:787–793, 1969.

[59] R. E. Showalter. Local regularity, boundary values and maximum principles for pseudoparabolic equations. *Applicable Anal.*, 16(3):235–241, 1983.

[60] R. E. Showalter and T. W. Ting. Pseudoparabolic partial differential equations. *SIAM J. Math. Anal.*, 1:1–26, 1970.

[61] Tsuan Wu Ting. Parabolic and pseudo-parabolic partial differential equations. *J. Math. Soc. Japan*, 21:440–453, 1969.

[62] J. Touma and M. Vauclin. Experimental and numerical analysis of two-phase infiltration in a partially saturated soil. *Transport in Porous Media*, pages 27–55, 1986.

[63] Hong Wang, Richard E. Ewing, and Thomas F. Russell. Eulerian-Lagrangian localized adjoint methods for convection-diffusion equations and their convergence analysis. *IMA J. Numer. Anal.*, 15(3):405–459, 1995.

[64] M. Weiler, T. Uchida, and J.J. McDonnell. Connectivity due to preferential flow controls water flow and solute transport at the hillslope scale. In D. Post, editor, *Proc. MODSIM 2003, Interactive modeling of Biophysical, Social and Biological Systems for Resource Management Solutions*, pages 398–403, 2003.

[65] Alan Weiser and Mary Fanett Wheeler. On convergence of block-centered finite differences for elliptic problems. *SIAM J. Numer. Anal.*, 25(2):351–375, 1988.

[66] D. Wildenschild, J. W. Hopmanns, and J. Simunek. Flow rate dependence of soil hydraulic characteristics. *Soil Sci. Soc. Am. J.*, 65:35–48, 2001.

[67] D. Wildenschild, J.W. Hopmans, A.J.R. Kent, and M.L. Rivers. A quantitative study of flow-rate dependent processes using x-ray microtomography. 2004.

[68] D. Wildenschild and K.H. Jensen. Laboratory investigations of effective flow behavior in unsaturated heterogeneous soils. *Water Res. Research*, 35(1):17–27, January 1999.

[69] Carol S. Woodward and Clint N. Dawson. Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media. *SIAM J. Numer. Anal.*, 37(3):701–724 (electronic), 2000.

Department of Mathematics, Oregon State University, Corvallis, OR 97331
E-mail: mpesz@math.oregonstate.edu and yis@math.oregonstate.edu
URL: http://www.math.oregonstate.edu/∼mpesz/ and
URL: http://www.math.oregonstate.edu/people/view/yis

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 150–156

HOMOGENIZATION OF SECONDARY-FLUX MODELS OF PARTIALLY FISSURED MEDIA

MALTE A. PETER AND RALPH E. SHOWALTER

Abstract. Fully-saturated and partially fissured media, in which supplementary flow and transport arise from direct cell-to-cell diffusion paths, have been described accurately over a wide range of scales by discrete *secondary-flux models*. These models were constructed as an extension of classical double-porosity models for totally fissured media by two-scale modeling considerations. There is some substantial literature on the analysis of continuously distributed secondary-flux models, and the corresponding discrete models have been proven to give efficient and accurate simulations when compared to recently available experimental data. These are particularly effective in the presence of advection. In this note, a summary description is given for the two-scale convergence of the discrete secondary-flux model to the corresponding continuous double-porosity secondary-flux model.

Key Words. secondary-flux, partially fissured porous media, homogenization, multiscale flow and transport.

1. Introduction

Problems of flow and transport through porous media lead to initial–boundary-value problems for a coupled elliptic–parabolic system of partial differential equations of elliptic and parabolic type. The fluid flow is described by an elliptic equation, and its solution provides the velocity for a parabolic equation with advection for the concentration u of a dissolved chemical transported by that flow. When the process takes place in a non-homogeneous medium, the coefficients vary on such a small scale that computation of the solution is very intensive and an upscaled model is needed. We shall consider the generic case of the single parabolic equation in a periodic medium of very small period $\varepsilon > 0$. This provides an indication of the corresponding results for the full system of flow and transport.

The locally representative unit cell is given in the two parts, $Y = Z^{\mathrm{f}} \cup Z^{\mathrm{s}}$, and then it is scaled to εY in the ε-periodic structure. In the classical case of the diffusion equation for transport, the diffusion coefficient varies between two constants, D^{f} on the *fast region* Z^{f} and D^{s} on the *slow* region Z^{s} of the unit cell Y. We denote the fine-scale coefficient in this situation by $D_\varepsilon(x) = [D^{\mathrm{f}}, D^{\mathrm{s}}; \varepsilon]$. The system is homogenized by taking the two-scale limit as $\varepsilon \to 0$, and the limit of its solution $u_\varepsilon(x,t)$ is the solution $u(x,t)$ of an equation of the same form but with the constant effective coefficient $\tilde{D}$. The formulae for $\tilde{D}$ show that the fast and slow regions are *flux coupled* through the gradient of the solution on the two regions. The gains of this homogenized model are that the fine-scale geometry is averaged out, so it is computationally straightforward, and it provides a good approximation of the real situation in the low-contrast cases when ε is small. See [7]

Received by the editors February 1, 2008 and, in revised form, March 1, 2008.
2000 *Mathematics Subject Classification.* 76S05, 35B27, 74Q15, 35R10.

for detailed expositions of various approaches and background in homogenization of porous media.

However, such models do not recover the tailing effects that are observed in experiments or in simulations when the contrast $D^{\mathrm{f}}/D^{\mathrm{s}}$ is large, for then there are consequential memory effects due to the relatively slower release of the solute stored in the small cells. A very special situation is the *obstacle problem* which corresponds to the extreme case of $D^{\mathrm{s}} = 0$. We denote the corresponding effective coefficient by $\tilde{D}^0$. Here, of course, there are no such memory effects, as there is no secondary storage, and this situation is described well by the preceding classical case. It is the cases of intermediate contrast that require better modeling.

The situation of highly-heterogeneous media in which the contrast between fast and slow regions is *very high* can be described as above but with the diffusion coefficient $D_\varepsilon(x) = [D^{\mathrm{f}}, \varepsilon^2 D^{\mathrm{s}}; \varepsilon]$ scaled as indicated in the slow region. Here the contrast is balanced with the cell size to maintain the two-way coupling of concentration and flux between the slow cells and the fast surrounding region. The limit leads to a *system* whose structure is quite different from the original single equation, namely, a macro-equation for an unknown $u(x,t)$ given on the macroscopic medium and a family of micro-equations for unknowns $U(x,y,t)$ given in the local reference cell at each point x of the macroscopic region. The cell solution provides the *source term* or input $q(x,t) = \int_{\partial Z^{\mathrm{s}}} D^{\mathrm{s}} \nabla_y U \cdot \nu \, \mathrm{d}\sigma$ back into the macro-equation, while the macro-variable enters the cell problem through the boundary condition

$$U(x,y,t) = u(x,t), \quad y \in \partial Z^{\mathrm{s}}. \tag{1}$$

This is the *double-porosity model* of Arbogast, Douglas & Hornung [2]. It is a large fully-coupled system, with a local diffusion problem at each point in the medium, but the structure is highly parallel and amenable to computation. It is *value* or *concentration coupled* into the cells and *gradient* or *flux coupled* into the macro-equation. The gain of this model includes the additional *secondary-storage* via the coupling of the fast and slow components and some of the resultant tailing effects and memory effects observed in experiments but unattainable with the classical model. The assumptions depend on the critical contrast ε^2 between coefficients. It was observed in [9] that the coefficients in the macro-equation are precisely those of the corresponding obstacle problem.

The double-porosity model completely misses any advective effects at the cell level, since the input to the cell (1) is constant on the local boundary. In order to couple the cells more tightly to the surrounding medium, the boundary condition (1) was replaced with the *affine* constraint

$$U(x,y,t) = u(x,t) + \nabla u(x,t) \cdot (y - y_0), \quad y \in \partial Z^{\mathrm{s}}, \tag{2}$$

by Peszyńska & Showalter [9]. Their objective was to include the local advective contributions and accurately model the full range of contrasts that were reported in the extensive experiments [13]. They showed the source term $q(x,t)$ needs to be altered to maintain conservation of mass, and this leads to the *secondary-flux* term. With the affine coupling into the cells, this model captures advection effects and contributes both the secondary-storage and the secondary-flux which are added back through the source term to the macro-equation. With this tighter coupling through both values and gradients, this model can cover a wide range of contrasts and accurately reproduce the break-through curves throughout the entire range of contrasts. See [9] for further discussion.

Since $D_\varepsilon = [D^{\mathrm{f}}, D^{\mathrm{s}}; \varepsilon]$ suits the very-low-contrast case while $D_\varepsilon = [D^{\mathrm{f}}, \varepsilon^2 D^{\mathrm{s}}; \varepsilon]$ describes the very-high-contrast case well, it is tempting to expect that an intermediate choice, *e.g.*, $D_\varepsilon = [D^{\mathrm{f}}, \varepsilon D^{\mathrm{s}}; \varepsilon]$, might be appropriate for the intermediate contrast. This is untrue, however, because, in this intermediate-contrast situation, advective effects become important [13]. These can not be captured by microscopic models describing diffusive transport only, cf. [11, 10]. The cell variable U needs to *see* the gradient of the macro-variable, ∇u, in order to account for an additional advection. This is accomplished by a condition of the type (2).

Discrete models of secondary-flux type were recently introduced in [9], although the continuous analogues were developed in [5] and [6] without any justification. The affine constraint (2) had been introduced earlier in the numerical work of Arbogast [3]. His *viscous* dual-porosity model used this pressure gradient to substantially improve the simulations of recovery at later times. These had been inadequate with only the pressure coupling of (1), and it was recognized that some modification of the source term $q(x,t)$ in the macro-equation was needed. We shall show here that the continuous model is not *ad hoc*, but is obtained as the two-scale limit of the corresponding exact microproblem of discrete type introduced in [9]. The detailed analysis of the problem will be presented in a forthcoming publication.

2. Statement of the problem

Let $Y = (0,1)^n$ be the reference cell, made up of two distinct parts Z^{f} and Z^{s} where $\overline{Z}^{\mathrm{s}} \subset Y$, and let $\Gamma = \partial Z^{\mathrm{s}}$. Given the open bounded Lipschitz-domain $\Omega \in \mathbb{R}^n$ and $\varepsilon > 0$, we define $\Omega_\varepsilon^\alpha = \Omega \cap \operatorname{int} \bigcup_k \varepsilon \overline{Z}_k^\alpha$, $\alpha \in \{\mathrm{f}, \mathrm{s}\}$, where the subscript k denotes translation of the set by the n-tuple of integers $k \in Z^n$. Similarly, $\Gamma_\varepsilon = \Omega \cap \bigcup_k \varepsilon \Gamma_k$. The time interval under consideration is $S = (0,T)$ where $T > 0$.

Let $[x]_Y$ denote the unique integer combination $\sum_{i=1}^n k_i e_i$ of the periods such that $\{x\}_Y = x - [x]_Y$ belongs to $[0,1)^n$. The vector e_j is the jth unit vector in n-dimensional Euclidean space. Note that we have $x = \varepsilon([x/\varepsilon]_Y + \{x/\varepsilon\}_Y)$ for any $x \in \mathbb{R}^n$.

Let $\mathcal{T}_\varepsilon \colon L^p(\Omega) \to L^p(\Omega \times Y)$, $p \in [1, \infty]$, be the periodic unfolding operator [4], i.e. for $u \in L^p(\Omega)$, extended by zero outside of Ω, we define

$$\mathcal{T}_\varepsilon(u)(x,y) = u(\varepsilon[x/\varepsilon]_Y + \varepsilon y) \text{ for } x \in \Omega \text{ and } y \in Y. \tag{3}$$

We define $z_\varepsilon(x)$ to be the function mapping each x to its part in εY translated by y_0,

$$z_\varepsilon(x) = (\{x/\varepsilon\}_Y - y_0), \tag{4}$$

where y_0 is the centroid of the interior region of the unit cell, Z^{s}. Moreover, we denote the mean value of a $u \in W^{1,2}(\Omega_\varepsilon^{\mathrm{f}})$ on each $\varepsilon\Gamma_k$ by $m_0(u)$,

$$m_0(u) = \frac{1}{|\Gamma|} \int_\Gamma \mathcal{T}_\varepsilon(u)(x,y)\, \mathrm{d}\sigma_y, \tag{5}$$

and the mean value of a vector-valued $v \in [L^2(\Omega_\varepsilon^{\mathrm{f}})]^n$ in each $\varepsilon Z_k^{\mathrm{f}}$ by $m_1(v)$,

$$m_1(v) = \frac{1}{|Z^{\mathrm{f}}|} \int_{Z^{\mathrm{f}}} \mathcal{T}_\varepsilon(v)(x,y)\, \mathrm{d}y. \tag{6}$$

Notice that m_0 and m_1 are constant in each cell εY_k.

All source terms are combined in the functions $f_\varepsilon^\alpha = f_\varepsilon^\alpha(x,t) = f^\alpha(x, x/\varepsilon, t)$, $\alpha \in \{\mathrm{f}, \mathrm{s}\}$, whose extensions by zero to all of Ω are assumed bounded independently of ε in $L^2(\Omega \times S)$. The coefficient functions $D_\varepsilon^\alpha = D_\varepsilon^\alpha(x,t) = D^\alpha(x, x/\varepsilon, t)$ are assumed bounded from above and away from zero by $D_0^\alpha > 0$ independently of

ε, $\alpha \in \{\mathrm{f},\mathrm{s}\}$, and they are supposed to be admissible test functions in two-scale-convergence sense [1].

The micro-problem under consideration is given by

$$\partial_t u_\varepsilon^{\mathrm{f}}(x,t) - \nabla\cdot(D_\varepsilon^{\mathrm{f}}(x,t)\nabla u_\varepsilon^{\mathrm{f}}(x,t)) = f_\varepsilon^{\mathrm{f}}(x,t), \quad x\in\Omega_\varepsilon^{\mathrm{f}},\ t\in S, \tag{7a}$$

$$\partial_t u_\varepsilon^{\mathrm{s}}(x,t) - \nabla\cdot(\varepsilon^2 D_\varepsilon^{\mathrm{s}}(x,t)\nabla u_\varepsilon^{\mathrm{s}}(x,t)) = f_\varepsilon^{\mathrm{s}}(x,t), \quad x\in\Omega_\varepsilon^{\mathrm{s}},\ t\in S, \tag{7b}$$

$$m_0(u_\varepsilon^{\mathrm{f}}) + \beta m_1(\nabla u_\varepsilon^{\mathrm{f}})\cdot z_\varepsilon = u_\varepsilon^{\mathrm{s}}(x,t), \quad x\in\Gamma_\varepsilon,\ t\in S, \tag{7c}$$

$$\begin{aligned} -D_\varepsilon^{\mathrm{f}}(x,t)\nabla u_\varepsilon^{\mathrm{f}}(x,t)\cdot\nu_\varepsilon^{\mathrm{f}} &= \varepsilon^2\frac{1}{|\Gamma|}\int_\Gamma \mathcal{T}_\varepsilon(D_\varepsilon^{\mathrm{s}}\nabla u_\varepsilon^{\mathrm{s}}\cdot\nu_\varepsilon^{\mathrm{s}})(x,y)\,\mathrm{d}\sigma_y \\ &\quad + \beta\varepsilon^2\frac{1}{|Z^{\mathrm{f}}|}\int_\Gamma \mathcal{T}_\varepsilon(D_\varepsilon^{\mathrm{s}}\nabla u_\varepsilon^{\mathrm{s}}\cdot\nu_\varepsilon^{\mathrm{s}} z_\varepsilon)(x,y)\,\mathrm{d}\sigma_y\cdot\nu_\varepsilon^{\mathrm{f}}, \quad x\in\Gamma_\varepsilon,\ t\in S. \end{aligned} \tag{7d}$$

$$u_\varepsilon^{\mathrm{f}}(x,t) = 0, \quad x\in\partial\Omega_\varepsilon^{\mathrm{f}}\cap\partial\Omega,\ t\in S, \tag{7e}$$

$$u_\varepsilon^{\mathrm{f}}(x,0) = u_0^{\mathrm{f}}, \quad u_\varepsilon^{\mathrm{s}}(x,0) = u_0^{\mathrm{s}}, \quad x\in\Omega. \tag{7f}$$

Note that, in particular, condition (7d) ensures that we have flux conservation across Γ_ε, since

$$\begin{aligned} -\int_{\Gamma_\varepsilon} D_\varepsilon^{\mathrm{f}}(x,t)\nabla u_\varepsilon^{\mathrm{f}}(x,t)\cdot\nu_\varepsilon^{\mathrm{f}}\,\mathrm{d}\sigma_x &= \varepsilon^2\frac{1}{|\Gamma|}\int_{\Gamma_\varepsilon}\int_\Gamma \mathcal{T}_\varepsilon(D_\varepsilon^{\mathrm{s}}\nabla u_\varepsilon^{\mathrm{s}}\cdot\nu_\varepsilon^{\mathrm{s}})(x,y)\,\mathrm{d}\sigma_y\,\mathrm{d}\sigma_x \\ &= \int_{\Gamma_\varepsilon}\varepsilon^2 D_\varepsilon^{\mathrm{s}}(x,t)\nabla u_\varepsilon^{\mathrm{s}}(x,t)\cdot\nu_\varepsilon^{\mathrm{s}}\,\mathrm{d}\sigma_x, \end{aligned} \tag{8}$$

where we have used the norm identity

$$\int_{\Gamma_\varepsilon} v(x)\,\mathrm{d}\sigma_x = \frac{1}{|\Gamma|}\int_{\Gamma_\varepsilon\times\Gamma}\mathcal{T}_\varepsilon(v)(x,y)\,\mathrm{d}\sigma_y\,\mathrm{d}\sigma_x. \tag{9}$$

For the weak formulation, the following function space is used,

$$\begin{aligned} \mathcal{V}_\varepsilon(\Omega) = \{(u_\varepsilon^{\mathrm{f}},u_\varepsilon^{\mathrm{s}}) &\in L^2(0,T;W^{1,2}(\Omega_\varepsilon^{\mathrm{f}}))\times L^2(0,T;W^{1,2}(\Omega_\varepsilon^{\mathrm{s}}))\,| \\ &u_\varepsilon^{\mathrm{f}} = 0 \text{ on } \partial\Omega_\varepsilon^{\mathrm{f}}\cap\partial\Omega \text{ and } m_0(u_\varepsilon^{\mathrm{f}}) + \beta m_1(\nabla u_\varepsilon^{\mathrm{f}})\cdot z_\varepsilon = u_\varepsilon^{\mathrm{s}} \text{ on } \Gamma_\varepsilon\}, \end{aligned} \tag{10}$$

and we write $u(t) = u(\,\cdot\,,t)$,

$$(u(t)\,|\,v(t))_{\Omega_\varepsilon^\alpha} = \int_{\Omega_\varepsilon^\alpha} u(x,t)v(x,t)\,\mathrm{d}x, \qquad (u\,|\,v)_{\Omega_\varepsilon^\alpha,t} = \int_0^t (u(s)\,|\,v(s))_{\Omega_\varepsilon^\alpha}\,\mathrm{d}s. \tag{11}$$

A weak form of problem (7) is defined as follows: find $(u_\varepsilon^{\mathrm{f}},u_\varepsilon^{\mathrm{s}})\in\mathcal{V}_\varepsilon(\Omega)$ with $(u_\varepsilon^{\mathrm{f}}(0),u_\varepsilon^{\mathrm{s}}(0)) = (u_0^{\mathrm{f}},u_0^{\mathrm{s}})$ such that

$$\begin{aligned} &(\partial_t u_\varepsilon^{\mathrm{f}}(t)\,|\,\phi^{\mathrm{f}}(t))_{\Omega_\varepsilon^{\mathrm{f}}} + (\partial_t u_\varepsilon^{\mathrm{s}}(t)\,|\,\phi^{\mathrm{s}}(t))_{\Omega_\varepsilon^{\mathrm{s}}} + (D_\varepsilon^{\mathrm{f}}(t)\nabla u_\varepsilon^{\mathrm{f}}(t)\,|\,\nabla\phi^{\mathrm{f}}(t))_{\Omega_\varepsilon^{\mathrm{f}}} \\ &\qquad + \varepsilon^2(D_\varepsilon^{\mathrm{s}}(t)\nabla u_\varepsilon^{\mathrm{s}}(t)\,|\,\nabla\phi^{\mathrm{s}}(t))_{\Omega_\varepsilon^{\mathrm{s}}} = (f_\varepsilon^{\mathrm{f}}(t)\,|\,\phi^{\mathrm{f}}(t))_{\Omega_\varepsilon^{\mathrm{f}}} + (f_\varepsilon^{\mathrm{s}}(t)\,|\,\phi^{\mathrm{s}}(t))_{\Omega_\varepsilon^{\mathrm{s}}} \end{aligned} \tag{12}$$

for all $(\phi^{\mathrm{f}},\phi^{\mathrm{s}})\in\mathcal{V}_\varepsilon(\Omega)$ and a.e. $t\in S$.

The following proposition ensures that (12) is an appropriate weak form of (7), the proof of which makes extensive use of (9) and the fact that for a function $\phi^{\mathrm{f}}\in C_0^\infty(\Omega_\varepsilon^{\mathrm{f}}\times S)$, we have

$$m_1(\nabla\phi^{\mathrm{f}}) = \frac{1}{|Z^{\mathrm{f}}|}\int_{Z^{\mathrm{f}}}\mathcal{T}_\varepsilon(\nabla\phi^{\mathrm{f}})(x,y)\,\mathrm{d}y = \frac{1}{|Z^{\mathrm{f}}|}\int_\Gamma \mathcal{T}_\varepsilon(\phi^{\mathrm{f}}\nu_\varepsilon^{\mathrm{f}})(x,y)\,\mathrm{d}\sigma_y. \tag{13}$$

Proposition 2.1. *Let $(u_\varepsilon^{\mathrm{f}},u_\varepsilon^{\mathrm{s}})\in\mathcal{V}_\varepsilon(\Omega)$ be a solution of (12). If the pair of functions $(u_\varepsilon^{\mathrm{f}},u_\varepsilon^{\mathrm{s}})$ also belongs to the space $C^1([0,T];C^2(\overline{\Omega}_\varepsilon^{\mathrm{f}}))\times C^1([0,T];C^2(\overline{\Omega}_\varepsilon^{\mathrm{s}}))$, it satisfies problem (7).*

Rather than the extra assumed smoothness of the solution, one can use the *abstract Green's theorem* [12, Proposition II.5.3] to characterize the strong form of the problem.

3. Macroscopic limit problems

We state the macroscopic limit problem satisfied by the limit functions of the sequences of solutions of (7) as $\varepsilon \to 0$. The limit functions of $u_\varepsilon^{\mathrm{f}}$ and $u_\varepsilon^{\mathrm{s}}$ are denoted by u^{f} and u^{s}, respectively. An outline of how these limit problems are obtained is given in §4.

The solution of a cell problem is required. Let ς_j, $j = 1, \dots, n$, be the Y-periodic solution of the cell problem

$$\begin{aligned} -\nabla_y \cdot (D^{\mathrm{f}}(x,y,t)(\nabla_y \varsigma_j(x,y,t) + e_j)) &= 0, \qquad y \in Z^{\mathrm{f}},\ x \in \Omega,\ t \in S, \\ -D^{\mathrm{f}}(x,y,t)(\nabla_y \varsigma_j(x,y,t) + e_j) \cdot \nu^{\mathrm{f}} &= 0, \qquad y \in \Gamma,\ x \in \Omega,\ t \in S, \end{aligned} \tag{14}$$

the weak form of which is given by

$$(D^{\mathrm{f}}(x,\,\cdot\,,t)(\nabla_y \varsigma_j(x,\,\cdot\,,t) + e_j) \,|\, \nabla_y \phi)_{Z^{\mathrm{f}}} = 0 \tag{15}$$

for all Y-periodic test functions ϕ. This allows the definition of the tensor $P^{\mathrm{f}} = [p_{ij}^{\mathrm{f}}]$ via

$$p_{ij}^{\mathrm{f}}(x,t) = \int_{Z^{\mathrm{f}}} D^{\mathrm{f}}(x,y,t)(\delta_{ij} + \partial_{y_i} \varsigma_j(x,y,t))\,\mathrm{d}y, \tag{16}$$

where δ_{ij} is the Kronecker delta. This turns out to be the effective tensor in the macroscopic limit problem. It is symmetric and positive definite, since D^{f} is bounded away from zero.

The limit problem reads as follows:

(17a)
$$\begin{aligned} &|Z^{\mathrm{f}}| \partial_t u^{\mathrm{f}}(x,t) - \nabla \cdot (P^{\mathrm{f}}(x,t) \nabla u^{\mathrm{f}}(x,t)) \\ &= \int_{Z^{\mathrm{f}}} f^{\mathrm{f}}(x,y,t)\,\mathrm{d}y - \int_\Gamma f^{\mathrm{int}}(x,y,t)\,\mathrm{d}\sigma_y + \nabla \cdot \int_\Gamma \beta f^{\mathrm{int}}(x,y,t)(y - y_0)\,\mathrm{d}\sigma_y,\ x \in \Omega,\ t \in S, \end{aligned}$$

$$u^{\mathrm{f}}(x,t) = 0,\ x \in \partial\Omega,\ t \in S, \tag{17b}$$

(17c)
$$\partial_t u^{\mathrm{s}}(x,y,t) - \nabla_y \cdot (D^{\mathrm{s}}(x,y,t) \nabla_y u^{\mathrm{s}}(x,y,t)) = f^{\mathrm{s}}(x,y,t),\ x \in \Omega,\ y \in Z^{\mathrm{s}},\ t \in S,$$

$$u^{\mathrm{f}}(x,t) + \beta \nabla u^{\mathrm{f}}(x,t) \cdot (y - y_0) = u^{\mathrm{s}}(x,y,t),\ x \in \Omega,\ y \in \Gamma,\ t \in S, \tag{17d}$$

$$u^{\mathrm{f}}(x,0) = u_0^{\mathrm{f}},\ u^{\mathrm{s}}(x,y,0) = u_0^{\mathrm{s}},\ x \in \Omega,\ y \in Z^{\mathrm{s}}. \tag{17e}$$

The weak form of problem (17) is: find $(u^{\mathrm{f}}, u^{\mathrm{s}}) \in L^2(0,T; W_0^{1,2}(\Omega)) \times [(u^{\mathrm{f}} + \beta \nabla u^{\mathrm{f}} \cdot (y - y_0)) + L^2(0,T; (L^2(\Omega); W_0^{1,2}(Z^{\mathrm{s}})))]$ with $(u^{\mathrm{f}}(0), u^{\mathrm{s}}(0)) = (u_0^{\mathrm{f}}, u_0^{\mathrm{s}})$ such that

(18a)
$$\begin{aligned} |Z^{\mathrm{f}}|(\partial_t u^{\mathrm{f}}(t) \,|\, \phi(t))_\Omega + (P^{\mathrm{f}}(t) \nabla u^{\mathrm{f}}(t) \,|\, \nabla \phi(t))_\Omega &= (\int_{Z^{\mathrm{f}}} f^{\mathrm{f}}(\,\cdot\,,y,t)\,\mathrm{d}y \,|\, \phi(t))_\Omega \\ &\quad - (\int_\Gamma f^{\mathrm{int}}(\,\cdot\,,y,t)\,\mathrm{d}y \,|\, \phi(t))_\Omega - (\int_\Gamma \beta f^{\mathrm{int}}(\,\cdot\,,y,t)(y - y_0)\,\mathrm{d}y \,|\, \nabla \phi(t))_\Omega, \end{aligned}$$

$$(\partial_t u^{\mathrm{s}}(t) \,|\, \psi(t))_{\Omega \times Z^{\mathrm{s}}} + (D^{\mathrm{s}}(t) \nabla_y u^{\mathrm{s}}(t) \,|\, \nabla_y \psi(t))_{\Omega \times Z^{\mathrm{s}}} = (f^{\mathrm{s}}(t) \,|\, \psi(t))_{\Omega \times Z^{\mathrm{s}}} \tag{18b}$$

for all $(\phi,\psi) \in L^2(0,T;W_0^{1,2}(\Omega)) \times L^2(0,T;(L^2(\Omega);W_0^{1,2}(Z^{\mathrm{s}})))$ and a.e. $t \in S$, where the interface term f^{int} is given by

$$f^{\mathrm{int}}(x,y,t) = D^{\mathrm{s}}(x,y,t)\nabla_y u^{\mathrm{s}}(x,y,t)\cdot\nu^{\mathrm{s}}, \quad x\in\Omega,\ y\in\Gamma,\ t\in S. \tag{19}$$

It is interesting to note that the limit problem would have been the same if we chose an average over Z^{f} instead of Γ in the definition of m_0 (cf. (5)).

4. Existence of solutions, a-priori estimates, convergence

We briefly summarize the steps of the homogenization analysis. The following theorem follows by standard techniques.

Theorem 4.1. *For fixed $\varepsilon > 0$, there exists a solution $(u_\varepsilon^{\mathrm{f}}, u_\varepsilon^{\mathrm{s}}) \in \mathcal{V}_\varepsilon(\Omega)$ of problem* (12) *such that*

$$|u_\varepsilon^{\mathrm{f}}(t)|_{\Omega_\varepsilon^{\mathrm{f}}} + |\nabla u_\varepsilon^{\mathrm{f}}|_{\Omega_\varepsilon^{\mathrm{f}},t} + |u_\varepsilon^{\mathrm{s}}(t)|_{\Omega_\varepsilon^{\mathrm{s}}} + \varepsilon|\nabla u_\varepsilon^{\mathrm{s}}|_{\Omega_\varepsilon^{\mathrm{s}},t} \le C, \tag{20}$$

for a.e. $t \in S$, where the constant C depends on T and the data but not on ε.

We use elements of two-scale convergence [8, 1] and periodic unfolding [4] in order to investigate the convergence of the sequences of solutions of (7) as $\varepsilon \to 0$. A key element in this analysis is the following proposition, which deals with the convergence of the terms $m_0(u_\varepsilon^{\mathrm{f}})$ and $m_1(\nabla u_\varepsilon^{\mathrm{f}})$.

Proposition 4.2. *The following convergence results hold:*

(a) $m_0(u_\varepsilon^{\mathrm{f}}) \longrightarrow u^{\mathrm{f}}$ *strongly in* $L^2(\Omega)$.
(b) $m_1(\nabla u_\varepsilon^{\mathrm{f}}) \longrightarrow \nabla u^{\mathrm{f}}$ *in two-scale sense.*

The limit problems associated with the limits of sequences of solution of the microproblem need to be identified. This is performed in two steps: the identification of the boundary condition (17c) and of equations (18a) and (18b). Particular attention needs to be paid to the recovery of the secondary-flux term in (18a). The main result can be summarized as follows:

Theorem 4.3. *The limit functions as $\varepsilon \to 0$ associated with a sequence of solutions of the microproblem* (12) *satisfy the macroproblem* (18).

Acknowledgments

The first author was supported by the German National Academic Foundation. A substantial amount of this research was undertaken during a visit of the first author to Oregon State University supported by a travel grant of the German Research Foundation (DFG) and the German Mathematical Society (DMV).
The second author was supported by the Department of Energy, Office of Science, through grants 98089 and 9001997.

References

[1] Allaire, G. Homogenization and two-scale convergence. *SIAM J. Math. Anal.* **23** (1992), 1482–1518.
[2] Arbogast, T., Douglas Jr., J. & Hornung, U. Derivation of the double porosity model of single phase flow via homogenization theory. *SIAM J. Math. Anal.* **21** (1990), 823–836.
[3] Arbogast, T. Computational aspects of dual-porosity models. In *Homogenization and porous media*, volume 6 of *Interdiscip. Appl. Math.*, pages 203–223. Springer, New York, 1997.
[4] Cioranescu, D., Damlamian, A. & Griso, G. Periodic unfolding and homogenization. *C. R. Acad. Sci. Paris, Ser. I* **335** (2002), 99–104.
[5] Clark, G. W. & Showalter, R. E. Fluid flow in a layered medium. *Quart. Appl. Math.* **52** (1994), 777–795.
[6] Cook, J. D. & Showalter, R. E. Microstructure diffusion models with secondary flux. *J. Math. Anal. Appl.* **189** (1995), 731–756.

[7] Hornung, U., editor, *Homogenization and porous media*, volume 6 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, 1997.
[8] Nguetseng, G. A general convergence result for a functional related to the theory of homogenization. *SIAM J. Math. Anal.* **20** (1989), 608–629.
[9] Peszyńska, M. & Showalter, R. E. Multiscale elliptic–parabolic systems for flow and transport. *Electron. J. Differential Equations* **2007** (2007), 1–30.
[10] Peter, M. A. Coupled reaction–diffusion processes and evolving microstructure: mathematical modelling and homogenisation. PhD dissertation, University of Bremen, 2006, also: Logos Verlag Berlin, 2007.
[11] Peter, M. A. & Böhm, M. Different choices of scaling in homogenization of diffusion and interfacial exchange in a porous medium. *Math. Meth. Appl. Sci.* 31 (11), p. 1257-1282.
[12] R. E. Showalter, *Monotone operators in Banach space and nonlinear partial differential equations*, volume 49 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1997.
[13] Zinn, B., Meigs, L. C., Harvey, C. F., Haggerty, R., Peplinski, W. J. & Freiherr von Schwerin, C. Environmental visualization of solute transport and mass transfer processes in two-dimensional conductivity fields with connected regions with high conductivity. *Environ. Sci. Technol.* **38** (2004), 3916–3926.

Department of Mathematics, University of Auckland, Private Bag 92019, Auckland 1142, NZ
E-mail: mpeter@math.auckland.ac.nz
URL: http://www.math.auckland.ac.nz/∼mpeter

Department of Mathematics, Oregon State University, Corvallis, OR, 97331-4605, USA
E-mail: show@math.oregonstate.edu
URL: http://www.math.oregonstate.edu/∼show

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 5, Supp , Pages 157–170

DOWNSCALING: A COMPLEMENT TO HOMOGENIZATION

ANNA TRYKOZKO, GEERT BROUWER, AND WOUTER ZIJL

Abstract. A groundwater flow model based on a specified hydraulic conductivity field in the modeling domain has a unique solution only if either the head or the normal flux component is specified on the boundary. On the other hand, specification of both head and flux as boundary conditions may be used to determine the conductivity field, or at least improve an initial estimate of it. The specified head and flux data may be obtained from measurements on the boundary, including the wells. We have presented a relatively simple, but instructive approach: the Double Constraint (DC) method. The method is exemplified in the context of upscaling and its inverse: downscaling. The DC method is not only instructive, but also easy to implement because it is based on existing groundwater modeling software. The exemplifications shown in this paper relate to downscaling and demonstrate that the DC method has practical relevance.

Key Words. Double Constraint Method, Downscaling, Inverse problems, Conductivity

1. Introduction

The Double Constraint (DC) method is a relatively simple, yet very instructive approach to inverse modeling. In this paper the DC method has been applied to downscaling, which can be considered as a practical complement to upscaling. However, the DC method is applicable in a wider range of settings, especially in applications in which wells play a role. The DC method is instructive, because it shows all the ingredients required for inverse modeling: measured heads and fluxes at the same location on the closed boundary, as well as estimated conductivities — the priors. At the same time, the method can be easily implemented, provided that groundwater modeling software is available.

In the context of groundwater flow, a forward model is a model in which the hydraulic conductivity is specified everywhere in the modeling domain. A forward model has a unique solution provided that appropriate boundary conditions are imposed. Considering groundwater flow this is the case only if either the head on a part of the boundary of the modeling domain, or the flux through that part of the boundary is specified in any point. Specification of both head and flux at that part of the boundary over-specifies the problem and has, therefore, no solution. However, such an over-specification may be used to improve the initially estimated conductivity field by conditioning it to the measured hydraulic data head and flux, in such a way that downscaling is meaningful. Determination of conductivities from additional boundary data is generally called inverse modeling. In our approach we follow the main steps of a method that has proved its applicability in Eletrical Impedance Tomography, [1, 4, 6, 9].

2000 *Mathematics Subject Classification.* 31B20, 35J25, 65C20.

After an introduction to downscaling in section 2, the double constraint method is presented in section 3. An exemplification of downscaling for a grid block far removed from wells is shown in section 4, where two isotropization equations — Wexler's equation and the square root equation — have been compared. A similar example is briefly presented in section 5. Section 6 presents a summary, conclusions and discussion, while section 7 shows the references.

For reasons of simplification, 2-dimensional problems will be considered. Extension to 3D problems is straightforward.

2. Downscaling

In this paper downscaling is considered as a practical complement to upscaling with application in groundwater flow modeling.

Upscaling starts with a fine-scale model with heterogeneous fine-scale conductivities in the elements (triangles, grid blocks) of a gridded rectangular upscaling cell. From these fine-scale conductivities the homogeneous effective coarse-scale conductivity of the upscaling cell is determined. A variety of upscaling methods has been applied and published, starting from well-known arithmetic and harmonic averages for flow respectively parallel and normal to layers, as well as the geometric average for fine-scale isotropic checkerboard patterns. For more complex fine-scale conductivity configurations the renormalization method can be applied, or a large class of methods based on fine-scale solution of the flow equation - see [7] for a review. Then based on specific discharge rates and head gradients in the fine-scale elements, the upscaled conductivity may be computed.

With respect to the latter class of methods, the question of boundary conditions to impose on the upscaling cell arises. Homogenization, probably the most popular method from this class, assumes periodicity of the porous medium and, as a consequence, periodic boundary conditions. Presumably, boundary conditions that are consistent with the actual flow might appear superior above the more-or-less arbitrarily chosen periodic boundary conditions. However, when using boundary conditions derived from an actual flow pattern, there is no consistency between different possible definitions of a large-scale conductivity, [11]. It should be mentioned that there exists another category of methods capable of dealing directly with a multiscale structure of the medium. A wide overview of such methods is given in [5]; this topic will not be further addressed in this paper.

A coarse-scale model consists of grid blocks (in a finite difference setting) in which each grid block has a coarse-scale conductivity that is obtained by upscaling from fine-scale conductivities. Once the solution of the flow problem in the large scale is computed, the modeler (the geohydrologist) may want to zoom in into the details of the groundwater flow in one or more coarse-scale grid blocks. If the original fine-scale conductivity distribution — from which the coarse-scale conductivity was derived by homogenization — is still known, we can run a fine-scale flow model on one large-scale cell with boundary conditions derived from the flow pattern calculated by the coarse-scale model. The fine-scale boundary conditions should be such that: (i) the total inflow through the boundary of the fine-scale model should be equal to the inflow calculated by the coarse-scale model, and (ii) the average head on each boundary node of the fine-scale model should be equal to the average head calculated by the coarse-scale model. Also wells may be considered as boundaries.

Since specification of two types of boundary conditions — both flux (discharge rate) and head — over-constrains a groundwater flow problem with specified conductivity distribution, we have to modify, or condition, the original fine-scale conductivities (the priors) in such a way that the conditioned conductivities honor both the flux and the head boundary conditions. The Double Constraint (DC) method presented here may be considered as a practical engineering approach to apply results obtained from simple homogenization methods to realistic, non-periodic media.

3. The Double Constraint method

The aim is to find conductivities that satisfy Darcy's law, the continuity equation, as well as both the flux and the head boundary conditions. Below we discuss the Double Constraint method, which presents, in a conceptually simple way, the ingredients of an inverse modeling technique: *both* measured boundary head *and* measured boundary flux complemented by prior conductivities in the flow domain. The DC method consists of three steps: (i) a head run with a forward model, (ii) a flux run with a forward model, and (iii) a post processing step. Simplicity is one of its great advantages: the method is based on standard finite difference or finite element groundwater flow models; therefore the main additional effort consists of implementing the post processing step. Our implementation is based on the finite element method.

Run 1: head constraining step. The original fine-scale conductivities (referred to as the priors) in the elements (triangles) of our finite element model are contained in system matrix A of the finite element model, while the specified heads (derived from the coarse-scale model) in the boundary nodes are contained in the right-hand side array B. Then the nodal heads X on the whole finite element mesh are calculated from solving the system of linear algebraic equations $AX = B$, from which the fluxes in the elements follow too. The calculated total flux through a boundary will generally differ from the total flux found by the coarse-scale model.

Run 2: flux constraining step. Now the boundary fluxes are specified resulting in the right-hand side array B. The original fine-scale conductivities (the priors), similarly as in the *run 1*, are now contained in system matrix A. The nodal heads X on the whole mesh are calculated from solving the system of linear algebraic equations $AX = B$, from which the fluxes in the elements follow. The calculated average head on a boundary will generally differ from the average head found by the coarse-scale model.

Post processing. In each element we determine the flux densities q'_x, q'_y obtained from flux-constraining *run 2*, as well as minus the head gradients $h_x = -\partial\phi/\partial x$, $h_y = -\partial\phi/\partial y$ obtained from head-constraining *run 1*. These fluxes and head gradients satisfy the measured flux and head boundary conditions, while the fluxes also satisfy the continuity equation (in discrete finite element form). To satisfy Darcy's law we define the conditioned conductivities (the *posteriors*) as $k_x = q'_x/h_x$ and $k_y = q'_y/h_y$. The thus-calculated conductivities are the fine-scale conductivities that belong to fine-scale flux densities q'_x, q'_y and heads ϕ in which we are interested.

Isotropization. If we prefer to avoid anisotropy, we define for each triangle an isotropic conductivity, either by Wexler's isotropization rule $k = -(q_x h_x + q_y h_y)/(h_x^2 + h_y^2)$, [6, 9] or by square root isotropization rule $k = \sqrt{k_x\, k_y}$, [3]. The above-described steps are repeated using the isotropized k's as priors until convergence to sufficient isotropy is obtained ($k_x/k_y \to 1$).

3.1. Wexler's method. Wexler's isotropization rule goes back to a class of methods generally referred to as Electrical Impedance Tomography (EIT), [1, 2, 4, 6, 9]. The term 'impedance' is taken from circuit theory where it denotes the ratio of the voltage (electric potential difference) across a circuit element to the electric current through that element. Originally, the method comes from medical imaging, where it is aimed at reconstruction of distribution of electric conductivity inside a human body, [4, 9]. As such, the Wexler's method falls into a broad category of inverse problems.

To give an argument for Wexler's isotropization formula we focus again on the two forward runs. Flux-constraining *run 2*, with Neumann boundary conditions, yielding the flux densities $q'_x = -k\,\partial\phi'/\partial x$, $q'_y = -k\,\partial\phi'/\partial y$ and head-constraining *run 1*, with Dirichlet boundary conditions, yielding the negative head gradients $h_x = -\partial\phi/\partial x$, $h_y = -\partial\phi/\partial y$. One cannot generally expect that $\mathbf{q}' + k\nabla\phi$ vanishes everywhere inside the region. As a consequence, a residual is obtained.

The optimization problem is then defined as minimization of the square of the residual over the computational domain; that is, $R = \int_\Omega (\mathbf{q}' + k\nabla\phi)(\mathbf{q}' + k\nabla\phi) d\Omega$ is minimal, where Ω denotes the modeling domain. In a finite element method $\mathbf{q}'$ and $\nabla\phi = -\mathbf{h}$ are defined element-wise, which means that the integral over Ω can be replaced with a summation of integrals over the elements (triangles) Ω_i: $R = \sum_i \int_{\Omega_i} (\mathbf{q}' + k_i\nabla\phi)\,(\mathbf{q}' + k_i\nabla\phi)\,d\Omega_i$.

Since conductivity k_i is assumed constant in each element, minimization of R by modifying the k_i requires $\delta R/\delta k_i = 2\int_{\Omega_i} (\mathbf{q}'\cdot\nabla\phi + k_i\nabla\phi\cdot\nabla\phi)\,d\Omega_i$ to be equal to zero for all elements Ω_i. Since in triangular elements $\mathbf{q}'$ and $\nabla\phi$ are constant, the integrand is constant yielding $\delta R/\delta k_i = 2\,(\mathbf{q}'\cdot\nabla\phi + k_i\nabla\phi\cdot\nabla\phi)\,A_i = 0$, where A_i is the surface area of triangle Ω_i. This results in the formula for the new conductivity value in each triangle: $k_i = -\mathbf{q}'\cdot\nabla\phi/(\,\nabla\phi\cdot\nabla\phi)$, which is the same as already given formula for Wexler's isotropization. An important distinction from other inversion procedures is that the error to be minimized by adjustment of the conductivity distribution is the difference between the interior current densities calculated from the Neumann and Dirichlet problems.

It should be remarked, however, that when choosing the residual to be minimized as $R' = \int_\Omega (\gamma\,\mathbf{q}' + \nabla\phi)\,(\gamma\,\mathbf{q}' + \nabla\phi)\,d\Omega$, where $\gamma = k^{-1}$ is the resistivity, we would end up with isotropization equation $\gamma_i = -\mathbf{q}'\cdot\nabla\phi/(\mathbf{q}'\cdot\mathbf{q}')$. In this case the error minimized by adjustment of the resistivity distribution is the difference between the interior potential differences calculated from the Neumann and Dirichlet problems.

The minimization approach presented above is also referred to as equation-error approach [2, 6]. The process is a least-square process which carries with it a measurement-error averaging property as well as stability. In [6] a slightly different formula for the updated k_i can be found: $k_i = \sqrt{\mathbf{q}'\cdot\mathbf{q}'/(\nabla\phi\cdot\nabla\phi)}$.

3.2. Differences with the EIT approach. Finally, it is important to emphasize the differences between our target and the aim originally addressed in the problems related to electrical impedance tomography (EIT). We are less restrictive in looking for the 'shape' of the conductivity distribution, in particular the problem of smoothed boundaries between areas of different conductivity, [2, 6], is not that crucial in downscaling groundwater flow patterns. Our main aim is to get a consistent formulation of a problem defined within the upscaling (downscaling) cell, with a simultaneous fulfillment of the two sets of boundary conditions. Also, the area of application has an influence on the quality of solution needed. On the other hand, in our case only one set of boundary conditions is applied, as opposed to

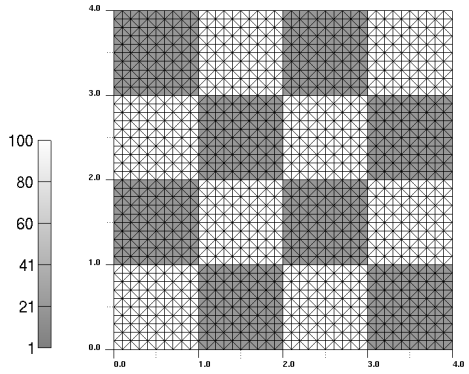


FIGURE 1. Homogenization square with checkerboard conductivity pattern.

classical EIT problems, where several measurements for different boundary fluxes and potentials are performed.

The approach presented in this paper has been successfully applied to match statistically generated permeability fields to data measured in wells, [3]. A square root method has been applied, with forward runs based on the finite difference flow model Modflow.

4. Example 1: Far field downscaling

4.1. Problem definition. In this section we present results obtained with the double constraint method. We consider a synthetic 2-dimensional case for which we apply respectively Wexler's and the square root isotropization. In the two constraining steps the flow equation is solved with a standard, conformal-nodal Finite Element Method with linear triangle-based elements. Conductivities are assumed constant within the triangles.

Let us consider a coarse-scale grid block that is far removed from a well. The square has a size of $[0, 4] \times [0, 4]$. The initial fine-scale conductivities within this square represent a checkerboard pattern with conductivities $k_1 = 1$ and $k_2 = 100$, as shown in Fig. 1. The discrete model is based on a grid with 1681 nodes, 4880 edges and 3200 triangles.

The aim is to condition the initial fine-scale conductivities in such a way that the two sets of boundary conditions dictated by the coarse model, one given as nodal boundary heads, the second given as boundary fluxes, are honored simultaneously.

The exact upscaled conductivity is equal to $\sqrt{k_1 \cdot k_2} = \sqrt{100 \cdot 1} = 10$, [8]. However, in case of a high contrast between k_1 and k_2 it is hard to obtain this value by discrete methods, such as finite elements or finite differences, [11]. If the homogenization procedure is performed for a square with the heterogeneity pattern shown in Fig. 1, even a relatively fine mesh consisting of 1681 nodes and 3200 triangles yields a computed effective permeability value equal to approximately 23 and is thus more than twice the exact value. This is important because the coarse-scale model acts on numerically upscaled conductivity values (here approximated as 23), whereas the fine-scale model exemplified here is based on exact conductivities (1 and 100 in checkerboard pattern). The great difference between the numerically

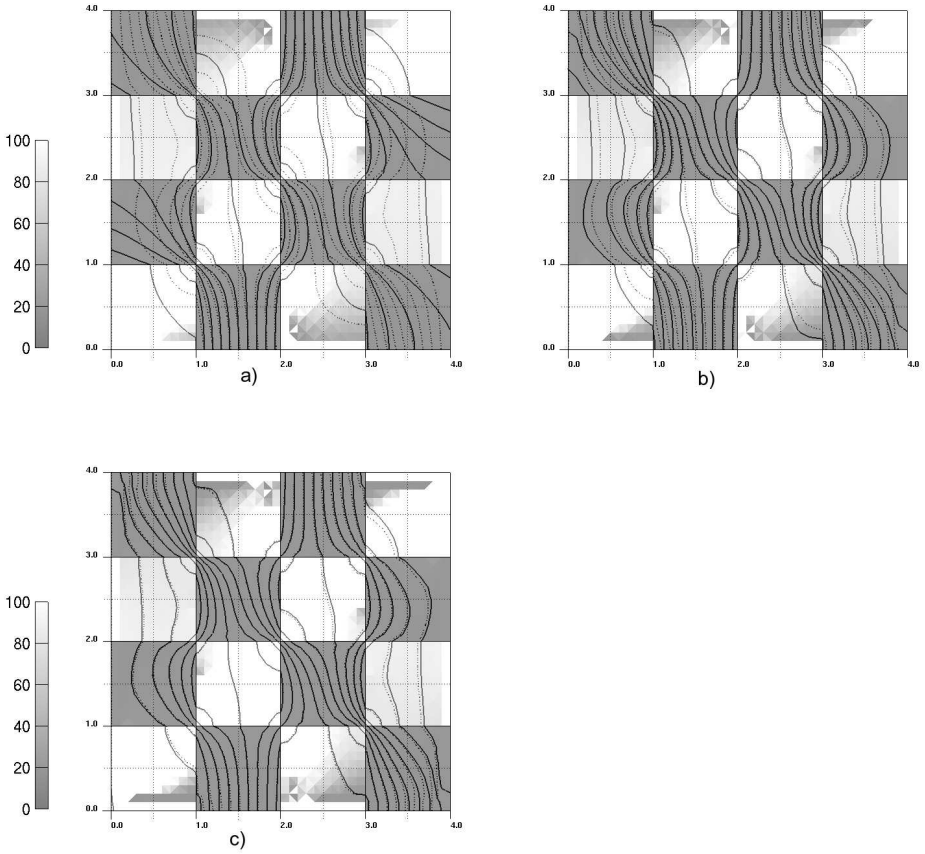


FIGURE 2. Solutions of the forward problems and conditioned conductivities. Solid lines: isolines of heads from *run 2* (specified boundary fluxes). Dashed lines: isolines of heads from *run 1* (specified boundary heads). a) after first iteration, b) after iteration 3, c) after iteration 8.

upscaled conductivity ($K = 23$) and the exact upscaled conductivity ($K = 10$) may have an additional influence on the discrepancy between the flux and/or potential boundary values of the coarse-scale and fine-scale model.

On the left and right sides of the domain fixed fluxes and fixed potential boundary conditions have been applied. On the horizontal sides the no flow boundary condition has been applied for the flux run and a linear difference in potential for the head run. These conditions are considered to come from a large-scale model.

The coarse-scale model yields heads in each node of the coarse-scale grid square, with a linear interpolation along the grid square's boundaries. The fine-scale boundary conditions do not follow that linear interpolation. They have been specified inversely proportional to the fine-scale conductivities of the triangles bordering at the boundaries, in such a way that the boundary averaged fine-scale head equals the boundary averaged coarse-scale heads. The coarse-scale model yields constant fluxes through the grid square's boundaries. Fine-scale boundary fluxes have not

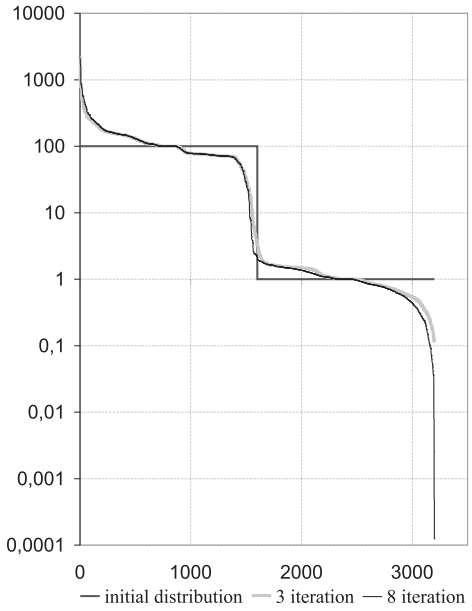


FIGURE 3. Distribution of conductivity values within the elements of the mesh during iterations. Far field case, Wexler's method.

been chosen equal to this constant flux. They have been specified proportional to the conductivities of the triangles bordering at the boundaries, in such a way that the total coarse-scale flux equals the total fine-scale flux through a boundary.

There is a question of compatibility of coarse-scale pressure and flux boundary conditions. If such compatibility is 'lacking', large changes in prior conductivity values occur during the iterations required for isotropization. A certain 'balance' among: flux conditions, head boundary condition and prior conductivities is established during iterations. This means that, if the two sets of boundary conditions are not 'equivalent' in quantity of flow, this will be obtained through generating changes in the conductivity field. On the other hand, if a consistent set of boundary conditions is given, then no change in priors occurs.

4.2. Wexler's isotropization. The results presented in Figs. 2 have been obtained with the DC method combined with Wexler's isotropization formula.

In Figs. 2 conductivities are visible only schematically. In order to get more insight in the values taken by the conductivities, Fig. 3 shows the distribution of the conductivities values obtained in the 3rd and 8th iteration.

A range of values taken by conditioned conductivity values may be considered as one of the characteristics of the performance of a method. It has been observed that there are a number of cells in which conductivities tend to relatively high and relatively small values. Conductivities generated after the 1st iteration were from the interval (0,32, 667), after the 3rd iteration from the interval (0,11, 2137), to become (1,24E-04, 2098) after the 8th iteration. The number of elements with such extreme values is not large, still this effect is highly undesirable because of the tendency of growing in subsequent iterations. Apart from this effect, the conductivities in the majority of elements does not change much in further iterations.

During the iterations it happens that negative values of the conductivity occur. This is mostly the case for regions with low conductivity and thus of a very slow

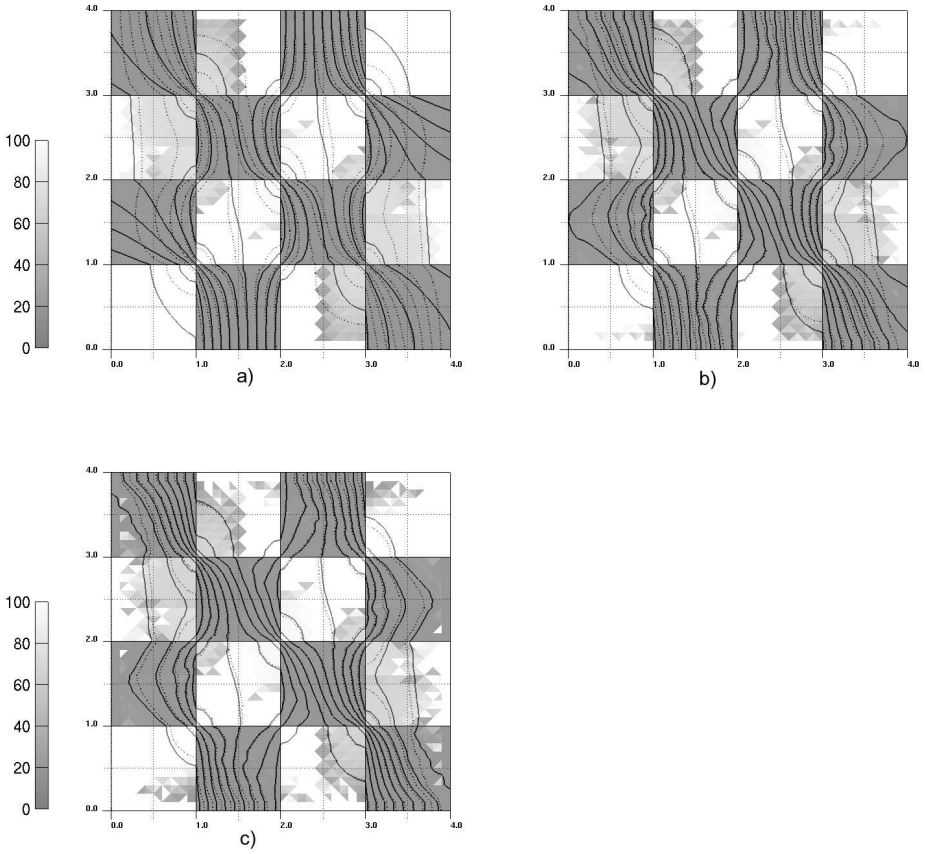


FIGURE 4. Solutions of the forward problems and conditioned conductivities. Solid lines: isolines of heads from *run 2* (specified boundary fluxes). Dashed lines: isolines of heads from *run 1* (specified boundary heads). a) after first iteration, b) after iteration 3, c) after iteration 8.

flow. In such a case the conductivity value obtained in a previous iteration was taken instead of the new prior. The number of negative values appearing at a given iteration step may be viewed as a measure of convergence of a method. In the computations performed for the case test presented in this section the number of negative conductivities varied from 20, during the 1st iteration, to 4 in the 8th iteration. It is important to note that in 2nd and 3rd iterations the number of negative conductivities was equal to zero. The deviation from isotropy cannot be measured in Wexler's method, but can in the square root method (Sec. 4.3).

4.3. Square root isotropization. Computations for the same initial conductivity pattern have been performed again, with a post-processing step based on square root isotropization. The results are presented in Figs. 4.

In the course of this method conductivities in x and y directions are computed independently. It is interesting to study the level of anisotropy created with the method. Fig. 5 gives a graphical comparison of anisotropy ratio obtained during

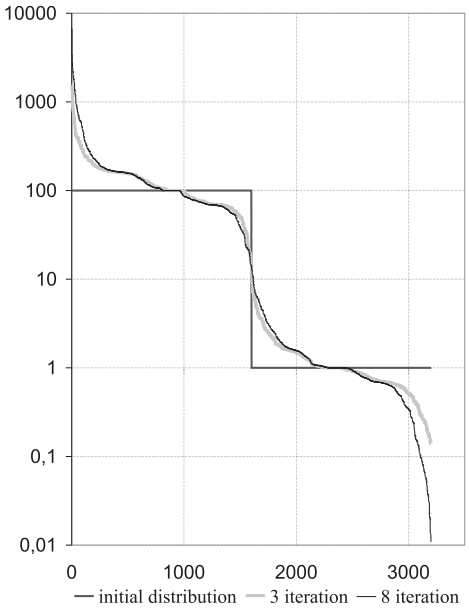


FIGURE 5. Distribution of conductivity values along elements of a mesh during iterations. Far field case, square root method.

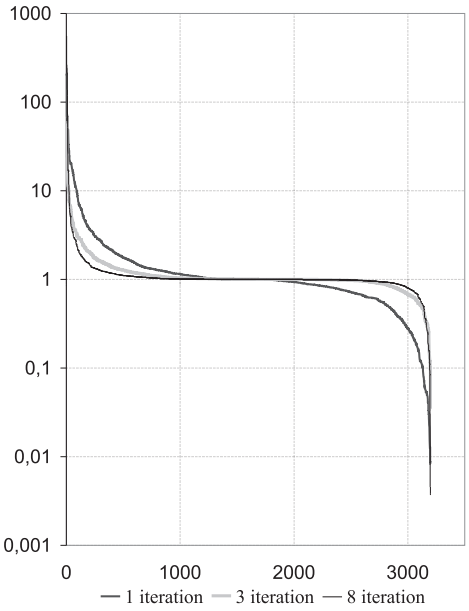


FIGURE 6. Anisotropy ratio during iterations. Far field case, square root method.

the 1st, 3rd and 8th iterations. A tendency of decreasing the number of elements with anisotropic conductivity has been observed, Fig. 6.

As compared to the Wexler's method, the square root method seems to be less stable in the sense of a tendency to create extremely high conductivity values.

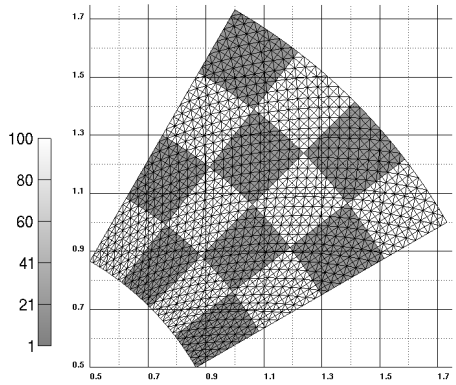


FIGURE 7. Picture in Cartesian plane of checkerboard conductivity pattern circular coordinate plane's square.

Moreover, special care must be taken while computing the conditioned conductivities $k_x = q'_x/h_x$ and $k_y = q'_y/h_y$ in regions of low conductivities (which results in low potential gradients).

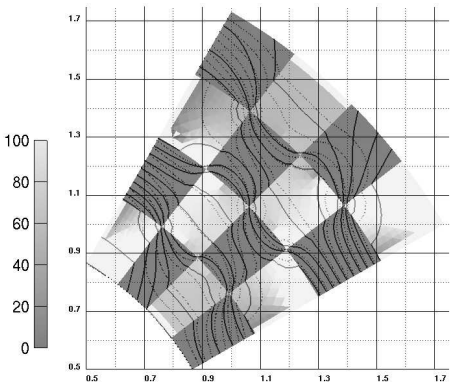
5. Example 2: Near well downscaling

In this section the Double Constraint method will be exemplified for a near-well problem. Let us consider a domain of a circular shape, as shown in Fig. 7. Although in a Cartesian x, y coordinate plane this shape is not a square, it does again represent a square ($[1, 2]\times[\pi/6, 2\pi/6]$) in the r, ϕ coordinate plane of a circular coordinate system. In that case the apparent conductivities do not only contain the ordinary conductivities, but also the metrical factors (or scale factors) of the circular coordinate plane, [10].

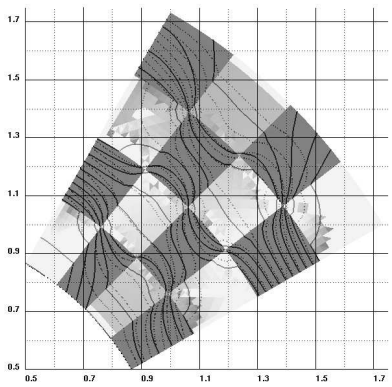
Results obtained with the Wexler's isotropization method will be presented together with results computed with the square root method, Figs. 8.

Differences in the distribution of conductivities obtained with the two methods visible in Figs. 8 reflect already mentioned 'smoothing' property of Wexler's isotropization method, as compared to the square root method. The conductivity distribution obtained during iterations for the Wexler's method are presented on Fig. 9, whereas Fig. 10 gives analogous results obtained for the square method. The anisotropy ratios obtained for the square root method are shown in Fig. 11.

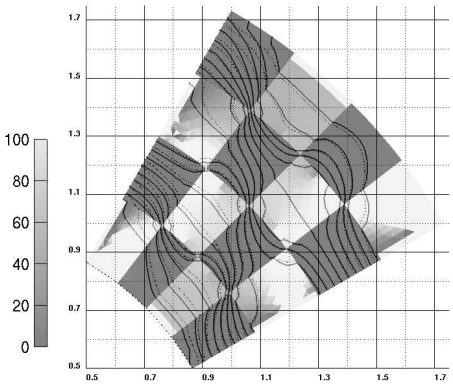
As in the former case, in a certain number of cells large and very small conductivities appeared. Conductivities generated after the 1st iteration were from the interval (0,20, 634), after the 3rd iteration from the interval (0,06, 1910), to become (4,69E-04, 4722) after the 8th iteration. Our observation is that once conductivity in an element becomes too large or too small, there is a tendency of unlimited growth or decrease. Apart from this effect, conductivities in the majority of the elements does not change much in further iterations. A very small number of negative values of conductivity occurred. The number of negative values appearing at a given iteration step may be viewed as a measure of convergence of the method. In the computations performed for the case test presented in this section the number of negative conductivities varied from 10, during the 1st iteration, to 3 in the 8th iteration.



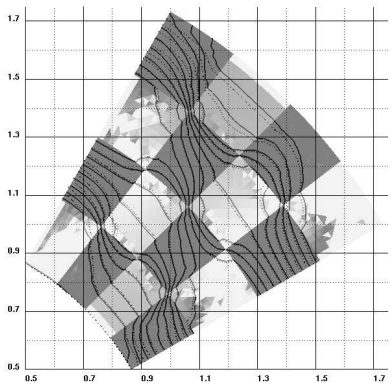
a) Wexler's method



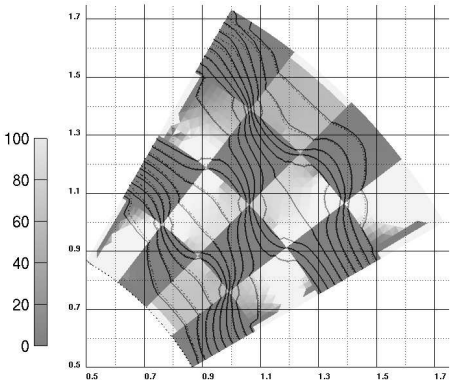
Square root method



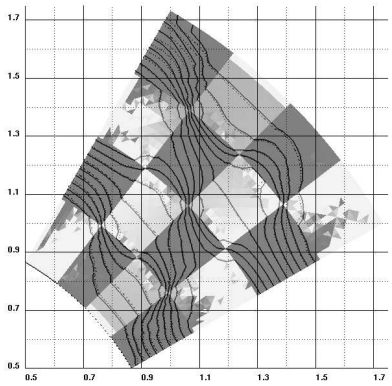
b) Wexler's method



Square root method



c) Wexler's method



Square root method

FIGURE 8. Solutions of the forward problems and conditioned conductivities. Solid lines: isolines of heads from *run 2* (specified boundary fluxes). Dashed lines: isolines of heads from *run 1* (specified boundary heads). a) after first iteration, b) after iteration 3, c) after iteration 8.

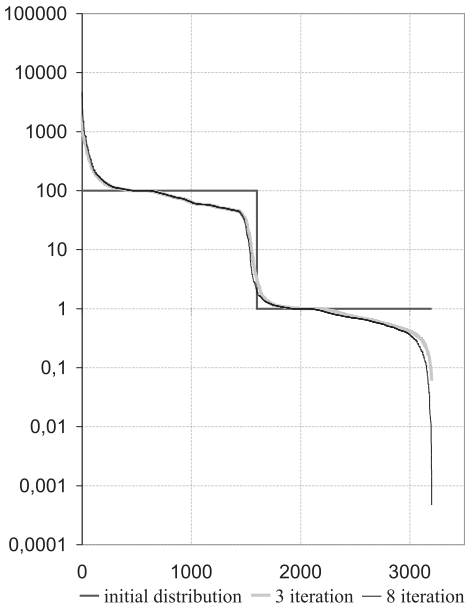


FIGURE 9. Distribution of conductivity values within elements of the mesh during iterations. Near well case, Wexler's method.

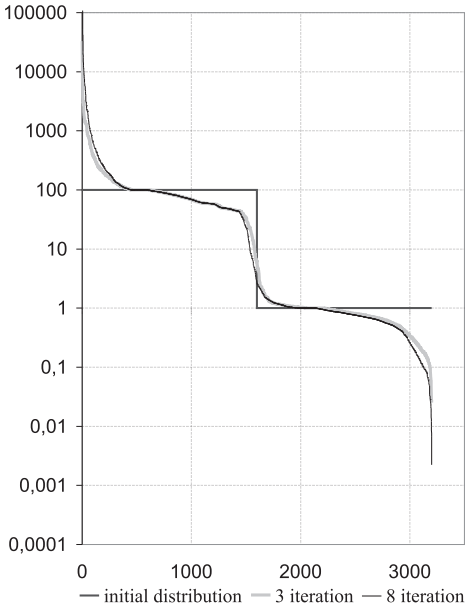


FIGURE 10. Distribution of conductivity values within elements of the mesh during iterations. Near well case, square root method.

6. Summary, conclusions and discussion

In the context of groundwater flow, a forward model is based on a specified hydraulic conductivity field in the modeling domain. A forward model has a unique solution only if, on a part of the modeling domain's boundary, either the head or

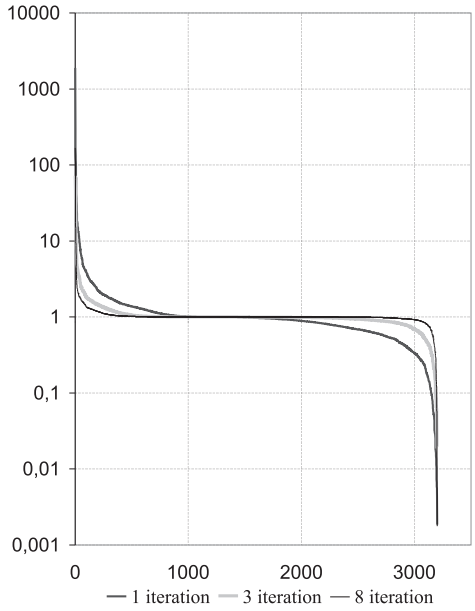


FIGURE 11. Anisotropy ratio during iterations. Near well case, square root method.

the normal flux component is specified. Specification of both head and flux at that part of the boundary over-specifies the problem and has, therefore, no solution. However, such an over-specification may be used to improve the initially estimated conductivity field by conditioning it to the measured hydraulic head and flux data. Determination of conductivities from such additional boundary data is generally called inverse modeling.

Here we have presented a relatively simple, but very instructive approach to inverse modeling, the Double Constraint (DC) method, in the context of upscaling and its inverse: downscaling. In the DC method a well is considered as a boundary and an observation well is a well with exactly zero flow rate and measured head. The DC method is instructive, because it shows all the ingredients required for inverse modeling: measured heads and fluxes at the same parts of the closed boundary, as well as prior knowledge on conductivities. In addition, the method can be implemented easily, because it is based on existing groundwater modeling software.

From the exemplification shown in this paper we observe that the DC method has practical relevance in the context of upscaling and downscaling. We observe that the DC method may be viewed as a kind of a smoothing procedure: large contrasts in conductivity are smoothed, still preserving the original flow pattern. We have also observed that a small percentage (less than 1%) of the conditioned conductivities is negative. Provided that the specified head at the inflow point of a stream tube is higher than the specified head at the outflow point of that stream tube, the effective resistance of that stream tube is positive. A negative resistivity (resistivity = 1/conductivity) means that most of the positive resistivities along the stream tube are too large, in such a way that one or a few negative resistances have to compensate in order to obtain the correct resistance (which is the weighted sum of the resistances along the stream tube). Therefore, if wanted, negative values

could be cured by "renormalization"; that is by making all positive resistances along the stream tube lower while making the negative resistivities along that tube positive, in such a way that the weighted sum of the resistivities yields the correct effective resistance. Such refinements will be the subject of a forthcoming paper.

References

[1] L. Borcea, Electrical impedance tomography, Inverse Problems, 18 (2002) 99-136.
[2] L. Borcea, G.A. Gray and Y. Zhang Y, Variationally constrained numerical solution of electrical impedance tomography, Inverse Problems 19 (2003) 1159-1184.
[3] G.K. Brouwer, P.A. Fokker, F. Wilschut and W. Zijl, A direct inverse model to determine permeability fields from pressure and flow rate measurements, Mathematical Geosciences (accepted).
[4] M. Cheney, D. Isaacson and J. Newell, Electrical impedance tomography, SIAM Review 41 (1999) 85-101.
[5] V. Kippe, J. Aarnes and K.-A. Lie, A comparison of multiscale methods for elliptic problems in porous media flow, Computational Geosciences (2007) accepted for publication.
[6] R.V. Kohn and A. McKenney, Numerical implementation of a variational method for electrical impedance tomography, Inverse Problems 6 (1990) 389-414.
[7] Ph. Renard and G. de Marsily, Calculating equivalent permeability: a review, Advances in Water Resources 20 (1997) 253-278.
[8] J.E. Warren and H.S. Price, Flow in heterogeneous porous media, SPE J, Sept (1961) 153-169.
[9] A. Wexler A., Electrical impedance imaging in two and three dimensions, Clin. Phys. Physiol. Meas., 9, Suppl. A, (1988) 29-33.
[10] W. Zijl and A. Trykozko, Numerical Homogenization of the Absolute Permeability Tensor Around a Well. SPE J, Dec. (2001) 399-408.
[11] W. Zijl and A. Trykozko, Numerical homogenization of the absolute permeability using the conformal-nodal and mixed-hybrid finite element method. Transport in Porous Media 44 (2001) 33-62.

ICM Interdisciplinary Centre of Mathematical and Computational Modelling, University of Warsaw, Pawinskiego 5a, PL-02-106 Warsaw, Poland
E-mail: A.Trykozko@icm.edu.pl

TNO Construction and Built Environment, P.O. Box 80015, NL-3508 TA Utrecht, The Netherlands
E-mail: geert.brouwer@tno.nl

Z-consult, c/o Laboratory of Hydrology, Vrije Universiteit Brussel (VUB), Pleinlaan 2, B-1050 Brussels, Belgium
E-mail: z-consult@zijl.be