

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

RICE UNIVERSITY

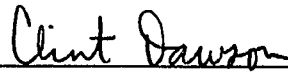
**Mixed Finite Element Methods for Variably
Saturated Subsurface Flow**

by

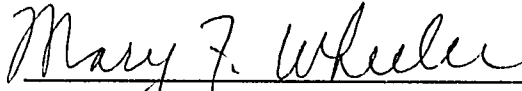
Carol A. San Soucie

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Doctor of Philosophy

APPROVED, THESIS COMMITTEE:



Clint N. Dawson, Co-Chairman
Associate Professor of Aerospace
Engineering and Engineering Mechanics
University of Texas at Austin



Mary F. Wheeler, Co-Chairman
Ernest and Virginia Cockrell Chair in
Engineering
University of Texas at Austin



Dan C. Sorensen
Professor of Computational and Applied
Mathematics



J. Ed Akin
Professor of Mechanical Engineering and
Materials Science

Houston, Texas

April, 1996

UMI Number: 9631088

UMI Microform 9631088
Copyright 1996, by UMI Company. All rights reserved.
This microform edition is protected against unauthorized
copying under Title 17, United States Code.

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Mixed Finite Element Methods for Variably Saturated Subsurface Flow

Carol A. San Soucie

Abstract

The flow of water through variably saturated subsurface media is commonly modeled by Richards' equation, a nonlinear and possibly degenerate partial differential equation. Due to the nonlinearities, this equation is difficult to solve analytically and the literature reveals dozens of papers devoted to finding numerical solutions. However, the literature also reveals a lack of two important research topics. First, no *a priori* error analysis exists for one of the discretization schemes most often used in discretizing Richards' equation, cell-centered finite differences. The expanded mixed finite element method reduces to cell-centered finite differences for the case of the lowest-order discrete space and certain quadrature rules. Expanded mixed methods are useful because this simplification occurs even for the case of a full coefficient tensor. There has been no analysis of expanded mixed methods applied to Richards' equation. Second, no results from parallel computer codes have been published. With parallel computer technology, larger and more computationally intensive problems can be solved. However, in order to get good performance from these machines, programs must be designed specifically to take advantage of the parallelism. We present an analysis of the mixed finite element applied to Richards' equation accounting for the two types of degeneracies that can arise. We also consider and analyze a two-level method for handling some of the nonlinearities in the equation. Lastly, we present results from a parallel Richards' equation solve code that uses the expanded mixed method for discretization.

Acknowledgments

I want to thank the Texaco Graduate Fellowship program, the Center for Subsurface Modeling Industrial Affiliates and the United States Department of Energy for supporting this work.

I wish to thank my two advisors, Clint Dawson and Mary Wheeler for their advice and help over the years. Clint Dawson has put much time and effort into my instruction and for his encouragement and energy, I am especially grateful. Mary Wheeler has been a strong source of support and inspiration for which I will always be appreciative.

I would also like to thank Lawrence Cowsar and Fredrik Sääf who have been the best of “brothers” the last few years.

Furthermore, I thank Laurie Feinswog, Cliff Nolan and the basement “dwellers” at Rice for their friendship. The time we have all shared has made my years in school memorable.

I would never have attended graduate school were it not for the encouragement and preparation given me by high school and college mentors, especially Paul Murrin and Martha Talbert, formerly of the Louisiana School for Math, Science and the Arts, and Elizabeth Swoope and Guillermo Ferreyra of Louisiana State University.

I owe much gratitude to Leonard Gray who first introduced me to research and set in motion events which have changed my life.

My mother, Dory San Soucie, my father, William San Soucie, and my brother, Paul San Soucie, have always been strong sources of encouragement and love. For that I can never fully express my appreciation.

Most importantly, I thank Robert Woodward for his never-ending words of encouragement, for his friendship and love and for the joy and happiness that he brings to life.

Contents

Abstract	ii
Acknowledgments	iii
List of Illustrations	vi
List of Tables	vii
1 Introduction	1
1.1 Introductory Remarks	1
1.2 Previous Work	1
1.3 Present Work	6
2 Physical Background	9
3 Discretization	17
3.1 Notation	17
3.2 Variational Formulations	18
3.3 Approximating Spaces	19
3.4 Time Discretization	23
4 An <i>a priori</i> Error Analysis of Richards' Equation	25
4.1 Partially to Fully Saturated Flow	26
4.2 Strictly Partially Saturated Flow	39
4.3 Unsaturated to Fully Saturated Flow	43
5 Two-Level Methods for Nonlinear Parabolic Equations	49
5.1 A Two-Level Finite Difference Scheme	50
5.1.1 A Coarse Grid Nonlinear Finite Difference Scheme	50
5.1.2 Fine Grid Linear Scheme	63
5.1.3 Extensions to Multiple Levels	67
5.2 A Two-Level Method for Richards' Equation	68

6	Implementation and Numerical Results	74
6.1	Implementation Issues	74
6.2	Numerical Results	77
6.2.1	A Known Solution Test Case	77
6.2.2	A One-Dimensional Flow Problem	79
6.2.3	A Three-Dimensional Irregular Geometry Flow Problem . . .	80
7	Conclusions	87
7.1	Summary	87
7.2	Future Work	88
	Bibliography	89

Illustrations

2.1	Typical van Genuchten curve of water content vs. matric potential. .	11
2.2	Typical van Genuchten curve of hydraulic conductivity vs. water content.	13
2.3	Typical van Genuchten curve of hydraulic conductivity vs. pressure head.	14
6.1	The 19-point discretization stencil.	75
6.2	Linear plot of convergence data.	78
6.3	Solutions for the one-dimensional test problem with various time step sizes.	81
6.4	Solutions for the one-dimensional test problem with various mesh sizes.	81
6.5	Three-dimensional single permeability layer test case after 45 simulation days.	83
6.6	Three-dimensional two permeability layer test case after 50 simulation days.	84
6.7	Three-dimensional irregular geometry test case after 5 simulation days.	85
6.8	Three-dimensional irregular geometry test case after 20 simulation days.	85
6.9	Three-dimensional irregular geometry test case after 50 simulation days.	86
6.10	Three-dimensional irregular geometry test case after 75 simulation days.	86

Tables

6.1	Convergence results for a three dimensional analytic test problem. . .	78
6.2	Physical data for the one-dimensional flow problem.	79
6.3	Physical data for the three-dimensional flow problem.	82

Chapter 1

Introduction

1.1 Introductory Remarks

Recent years have seen an increase in attention to modeling the flow of water through variably saturated porous media. This increase arises from heightened interest in finding appropriate sites for waste facilities and in evaluating the impact of current sites on local groundwater systems. One way to gain an understanding of the groundwater systems at these sites is through computer simulations of subsurface flow.

A commonly accepted mathematical model of water flow through variably saturated porous media is Richards' equation, a nonlinear parabolic partial differential equation well known in hydrology and related sciences. Richards' equation is expressed as,

$$\frac{\partial \theta(h)}{\partial t} + S_s \frac{\partial h}{\partial t} - \nabla \cdot K(h) \nabla h = f, \quad (1.1)$$

where h is the hydraulic head, θ is the moisture content of the soil, S_s is the specific storage of the medium, K is the hydraulic conductivity and f is a water source/sink term. The highly nonlinear nature of this equation makes analytical solutions difficult to find, so this equation is most often solved numerically.

In this thesis we formulate efficient discretization schemes based on the mixed finite element method for the solution of Richards' equation, prove *a priori* error estimates for these methods and show results from a computer program developed for parallel platforms which solves Richards' equation.

1.2 Previous Work

Before presenting the results of this thesis, a brief summary of previous work on the analysis and numerical solution of Richards' equation is in order. We first consider reasons for choosing mixed finite element methods and some results pertaining to these methods. We then mention previous work analyzing Richards' equation, and lastly, we discuss the literature pertaining to numerical solutions of the equation.

Mixed methods are considered for this work because they conserve mass on a cell-by-cell basis. This conservation of mass means that at any time, the flux out of each cell is equal to the flux into the cell plus any source. Galerkin finite elements only guarantee a global conservation of mass, meaning that the flux out of the domain is equal to that into the domain plus any source. Since Richards' equation is actually a conservation equation, mixed methods in some sense require the solution to hold cell-by-cell.

Mixed finite element methods for linear elliptic problems have been well studied [17, 56, 9]. Element spaces have been developed for both two and three dimensions and for many different element shapes [54, 50, 14, 15, 16]. For these equations analysis has shown that if h is the maximal mesh spacing, then optimal convergence of the lowest order mixed method is $O(h)$ for both the scalar and velocity variables. Moreover, superconvergence of $O(h^2)$ has been shown for the pressure and velocity variables at certain points [49, 62, 29, 25].

In the case of linear elliptic equations, Russell and Wheeler [59] have shown that for the lowest order Raviart-Thomas-Nedelec [54, 50] spaces on rectangles and for a diagonal tensor K , the use of certain quadrature rules simplifies the mixed method into a cell-centered finite difference scheme with a 5 point stencil in two dimensions and a 7 point stencil in three dimensions. Weiser and Wheeler [62] showed that this simpler scheme retains the convergence and superconvergence rates of the original method for both the pressure and velocity.

Although much analysis has been done on mixed finite element methods, most of it assumes that K is a diagonal and invertible tensor. However, a full tensor can arise when computing "effective permeabilities" as in upscaling from fine to coarse data [26] or when mapping a rectangular grid into a logically rectangular grid [7]. When the tensor is full it is not possible to derive a finite difference scheme equivalent to the mixed method. Recently, methods have been developed to handle a full, possibly noninvertible tensor [8, 19, 44]. In particular, Arbogast, Wheeler and Yotov have analyzed the expanded mixed finite element method [8]. This method simultaneously approximates the pressure, its gradient and the flux. Arbogast, Wheeler and Yotov showed that for the lowest order Raviart-Thomas-Nedelec space on parallelepipeds, a cell-centered finite difference scheme results from this method. In certain discrete norms and for linear elliptic equations, this scheme exhibits superconvergence of $O(h^2)$ for the scalar variable and of $O(h^{3/2})$ for its gradient and flux. However, in the interior of the domain, they show $O(h^2)$ for the last two of these.

There has been relatively little analysis of the mixed method applied to nonlinear equations. Milner [48] developed a mixed method for the solution of two-dimensional second order quasi-linear elliptic equations. He was able to show existence and uniqueness of a solution to his scheme as well as optimal convergence. Dawson and Wheeler [22] in the course of analyzing a two-grid scheme for three-dimensional problems derived optimal order estimates for the expanded mixed method applied to the nonlinear heat equation.

Richards' equation is particularly difficult to analyze since it can be degenerate, i.e. the $K(h)$ term can be 0 as can the time derivative of θ . There has been some recent work on the analysis of degenerate parabolic equations. Rose [57] considered the porous medium equation, which admits solutions lacking the regularity of classical solutions. He developed continuous and discrete time Galerkin finite element approximations and derived estimates based on assumed rates of degeneracy. Nochetto and Verdi [51] also considered degenerate parabolic equations and developed linear Galerkin finite element schemes with error estimates. Arbogast, Wheeler and Zhang [6] made use of the Kirchhoff transformation in order to develop estimates of the mixed method applied to degenerate parabolic equations but they assume a linear time derivative term.

In [4], Arbogast developed error estimates for Galerkin finite elements applied to Richards' equation. He allowed for the time derivative of θ to be 0 but assumed $K > 0$. Arbogast, Obeyeskere and Wheeler [5] developed estimates for the Galerkin method applied to Richards' equation in the case that both the time derivative of θ and the hydraulic conductivity are not 0.

When considering previous numerical work on Richards' equation, it is helpful to be familiar with some common formulations of the equation. Different formulations of Richards' equation have various advantages and disadvantages depending on the physical situation and the numerical scheme. Various formulations are possible due to a constitutive relationship between pressure head and water content. Probably the most common expression of the equation is formulated in terms of the pressure head only,

$$(S_s + C(h)) \frac{\partial h}{\partial t} - \nabla \cdot K(h) \nabla h = f,$$

where $C(h) = \partial\theta/\partial h$ is the water capacity. While this formulation gives the solution as pressure head, due to the way the time derivative is expressed, numerical schemes based on this form tend to be nonconservative. Another form is based on the water

content,

$$\frac{\partial \theta}{\partial t} - \nabla \cdot D(\theta) \nabla \theta = f,$$

where $D(\theta) = K(\theta)/(\partial \theta / \partial h)$. This form is advantageous in that it is in conservative form. However, for saturated media, θ becomes constant, D approaches infinity, and this form is no longer applicable. Furthermore, θ is not continuous across interfaces separating layers of two different soils. The pressure head is continuous across these discontinuities which makes head-based methods better suited for modeling flow in layered soils. However, researchers have found that schemes for the head-based formulation produce large mass balance errors [18, 39]. The formulation in equation (1.1) is the mixed form. This formulation is also mass conserving and gives the solution in terms of pressure.

Many papers have been published discussing numerical solutions to Richards' equation. The most common approaches use a low-order finite difference or finite element method in space with backward Euler or Crank-Nicholson time discretization and Newton or Picard iteration for the nonlinearities. We now briefly describe some of this work.

Allen and Murphy [2] in the context of collocation methods and Celia, Bouloutas and Zarba [18] in the context of finite differences and finite elements have formulated the modified Picard method for handling nonlinearities in the mixed form of Richards' equation. This method applies Picard iteration to the nonlinearities in the hydraulic conductivity, but uses a first order Taylor expansion of θ about the previous value for the time derivative term. This expansion results in the same linear system as the standard head-based scheme except that the right hand side also contains the time derivative of θ at the previous iteration for the given time level. Numerical results show this term helps to preserve mass balance that is lost with the standard head-based schemes.

Much work on the solution of the head-based form of the equation has focused on developing mass-conservative schemes for this nonconservative form. In [47], Milly formulated a mass-conservative scheme by using an average value of the water capacity over each time step. This averaging reduced error associated with evaluation of the function at a fixed point which may or may not represent the behavior over the entire time step. Kirkland, Hills and Wierenga [43] employ an update for θ based on computed flux values. This new θ update removes the nonconservative nature of the head-based scheme and preserves mass balance. Rathfelder and Abriola [53]

developed mass-conservative numerical solutions of the head-based form with both finite elements and finite differences. They make use of a chord-slope approximation of the water capacity term, C . In the finite difference case, their scheme for the head-based form results in the same discrete system as the scheme of Celia, et.al. for the mixed form of the equation.

Hills, Porro, Hudson and Wierenga [39] developed a scheme for the θ -based form. They modeled the discontinuities of θ by adding an additional source term expressed as a jump in θ values across interfaces. Comparisons between their water content-based and head-based forms show that the θ -based scheme is far better at conserving mass and is less sensitive to time step size in dry conditions than the pressure-based form. They point out that the main disadvantage of the θ -based scheme is its inapplicability to saturated flow.

Some authors have considered the Kirchhoff transform to numerically handle degeneracies. Haverkamp and Vauclin [37] compared a finite difference solution of the Kirchhoff transformed equation with the head-based form. They found that the transformed equation gave more accurate results but required much more compute time due to the need for integrated values of the transformation. Ross and Bristow [58] have discussed the transformation in the case of discontinuous hydraulic conductivities. They apply the Kirchhoff transformation element-by-element, then couple the elements together through the continuous pressure head at element boundaries.

Some authors have looked at variable transformations to switch between saturated and unsaturated conditions thereby using the θ -based form in unsaturated regions and the head-based form in saturated regions. Kirkland, Hills and Wierenga [43] transform Richards' equation in terms of a single variable which is defined as water content in dry conditions and pressure head otherwise. With this transformation the scheme generates very little mass balance errors, and is stable over a wide range of conditions. However, they find degradation in accuracy near the interface between saturated and unsaturated regions. Forsyth, Wu and Pruess [33] developed a similar scheme in that they switch from head-based to θ -based schemes depending on water saturation values. Their scheme differs from Kirkland et.al. in that the change in variables is performed after the equation is discretized, as opposed to rewriting the original equation in terms of a new, more general variable. Forsyth et.al. switch between pressure and water content by substituting the appropriate variable in the equation for each grid point. Numerical results show a significant improvement in computational speed by using this variable substitution method instead of standard

head-based methods for dry conditions. The reason for this improvement is that they are able to take larger time steps when the domain is unsaturated.

Huyakorn, Thomas and Thompson [40] compared the Newton and Picard methods. For a Galerkin method applied to the head-based form of the equation, they have found that even though the Newton iterations are each slower than the Picard iterations, in general, significantly fewer Newton iterations are required for convergence.

One difficulty with solving Richards' equation numerically is that despite the fact that the equation is parabolic, steep water saturation fronts can occur when modeling flow of water into very dry media. The saturation fronts can be very difficult to simulate numerically and common methods such as Galerkin finite element methods can produce sharp nonphysical oscillations near these fronts. Forsyth and Kropinski [32] have given monotonicity conditions for a head-based scheme. If these conditions are met, the solution will be non-oscillatory near steep fronts. They further indicate that only upstream weighting [61] for the K term as opposed to central weighting will give nonoscillatory solutions. Abriola and Lang [1] have shown that adding more degrees of freedom by using a higher order method near the front gives more accuracy than if the same number of new unknowns were introduced solely by grid refinement.

1.3 Present Work

Until now, there have been no estimates of the mixed finite element method applied to Richards' equation. This thesis will present a number of estimates for the expanded mixed method applied to the equation. We first present a continuous time analysis for the case where $K > 0$ and the time derivative of θ may be zero. This is the case for partially to fully saturated flow. For this situation, bounds of the error in approximating θ and the negative gradient of hydraulic head are derived in terms of a Hölder continuity parameter. Furthermore, an optimal bound for the nonlinear form,

$$\left(\int_0^T (\theta(p) - \theta(P), p - P) dt \right)^{1/2},$$

where p is the hydraulic head, is derived. The bound is optimal since it is equal to the order of truncation error for approximation with the same degree polynomial as the approximating space used in the method. In addition, this nonlinear form is bounded below by the error in θ and above by the error in p .

Next, we consider the case where $K > 0$ and the time derivative of θ is strictly nonzero. This is the case of strictly partially saturated flow. For this situation,

optimal convergence of the hydraulic head and its negative gradient are shown for a fully discrete time scheme.

The third case considered occurs when the tensor coefficient is positive semi-definite, and $\partial\theta/\partial p \geq 0$. This is the case of unsaturated to fully saturated. For this possibly degenerate situation, the Kirchhoff transform,

$$R(p) = \int_0^p k(\theta(\wp))d\wp,$$

is used, where p is the hydraulic head and $k(\theta(p))$ is the relative permeability. As seen below, this transformation moves the nonlinearity from the K term to the gradient. In the situation when $K = 0$, the problem solution lacks enough regularity to formulate a variational problem involving the time derivative of θ , with trial functions in L^2 . Thus, we follow the technique of Arbogast, Wheeler and Zhang [6] and formulate an integrated in time scheme. The error estimates for the resulting scheme applied to Richards' equation are optimal in the sense that they reduce to approximation error.

Having analyzed the expanded mixed method applied to Richards' equation, we turn to methods of handling the nonlinearities at the level of discretization. The approach used is that of J. Xu [63, 64] and Dawson and Wheeler [22]. In these works, the discretization scheme is applied to the nonlinear equation on a coarse grid, and the equation is then linearized about the coarse grid solution on the fine grid. Xu analyzed this scheme for Galerkin methods applied to nonlinear elliptic equations, and Dawson and Wheeler analyzed the scheme for the expanded mixed method applied to the nonlinear heat equation. As a first step in applying this scheme to Richards' equation, we analyze the scheme for a superconvergent cell-centered finite difference method also applied to the nonlinear heat equation. Then the scheme for the expanded mixed method applied to Richards' equation is discussed.

Although much computational work has been done in finding efficient ways of solving Richards' equation, to this author's knowledge there have been no published results from a parallel computer code. Results are given from a parallel, three-dimensional Richards' equation code, PREQS. This code uses a cell-centered finite difference scheme equivalent to the expanded mixed method with quadrature. One point upstream weighting is used to more accurately model the moving fronts. Parallelism is achieved by spatially decomposing the domain into subdomains and assigning one subdomain to each processor, and extra unknowns are introduced along subdomain interfaces in order to reduce communication requirements.

Results are given from a variety of test cases. The first case is a nonlinear parabolic equation with a source term chosen to guarantee a specific solution. A three-dimensional convergence analysis is done which indicates a spatial rate of convergence of almost $O(h^2)$. The second test case is a one-dimensional Richards' equation problem from Celia, Bouloutas and Zarba [18]. Celia et.al. measure the mass balance ratio which is the total amount of water entering at the boundaries of the domain divided into the time rate of change in water mass. For a mass conserving numerical method, this ratio should always be unity. Celia et.al. report a ratio of 1 for a mixed formulation scheme and ratios significantly less than 1 for a head-based scheme. The PREQS code always gives a ratio of 1, indicating conservation of mass. Lastly, results are given for a three-dimensional full tensor Richards' equation case using the general geometry techniques of Arbogast, Wheeler and Yotov [7]. These results indicate that the code predicts reasonable solutions to flow problems on general domains.

The rest of this document is organized as follows. In the next chapter an overview of the physical flow problem and assumptions leading to Richards' equation are given. In chapter 3, notation and discretization schemes are introduced. We summarize the mixed and expanded mixed finite element methods as well as discuss the Raviart-Thomas-Nedelec approximating spaces. In the following chapter an *a priori* error analysis of the expanded mixed method applied to Richards' equation is presented. Chapter 5 discusses a novel two-level method for handling the nonlinearities in the equation and chapter 6 presents the parallel Richards' equation code and numerical results. Lastly, chapter 7 gives a brief summary of the thesis and indicates directions for future work.

Chapter 2

Physical Background

In this chapter we give a brief description of the physical laws that lead to Richards' equation for flow of water through variably saturated porous media. For information beyond that presented here, the reader is referred to the books of Bear [11, Chapter 9], Fetter [31, Chapter 4] and Freeze and Cherry [34, Chapter 2].

The physical situation we are modeling is that of water flowing into a porous medium filled with air and a small amount of water.

Water saturation measures the amount of water in the medium and is defined as the fraction of total pore space that is filled with water. The term “variably saturated” refers to the possibility that the water saturation, s , can vary between some residual water saturation, s_r , and s_s , a fully saturated medium. The medium is called unsaturated if water saturation is less than s_s and saturated otherwise. Water saturation is closely related to the volumetric water content of the soil, θ , which is the fraction of total volume that is filled with water. The relationship between θ and s is,

$$\theta = \phi s,$$

where ϕ is the porosity, or amount of pore space per unit volume of the medium, and s is the water saturation.

In unsaturated flow, the driving force of the flow is the matric potential, Ψ , which has units of Newtons per square meter (N/m^2). This potential is caused by surface tension creating a negative pressure on the pore water and is a function of the volumetric water content of the soil, θ .

In unsaturated media, the matric potential is negative and is equal to the negative of capillary pressure, P_c , also having units of N/m^2 . The capillary pressure is related to the pressures of the other phases by,

$$P_c = p_a - p_w,$$

where p_a and p_w are the air and water pressures, respectively. For the case of a single water phase flowing into a porous medium, it is assumed that the air phase pressure

remains constant at atmospheric pressure. Thus, the capillary pressure is no longer a function of p_a . In this case, we will consider the water pressure as a gage pressure, i.e. $p_w = p'_w + p_a$ where p'_w is the absolute water pressure. Thus, $-P_c = p_w$. Capillary pressure can also be experimentally measured as a function of water saturation. The resulting curves exhibit hysteresis; they are different depending on whether water is flowing into (imbibition) or out of (drainage) the medium. In the work considered here, hysteresis will be neglected. Due to the relationship between capillary pressure and water saturation, we can write

$$s = P_c^{-1}(p_a - p_w),$$

where p_a is constant. This relation shows s as a function of p_w .

Van Genuchten [38] derived an empirical formula for the water content as a function of matric potential. This relationship is,

$$\theta = \theta_r + \frac{\theta_s - \theta_r}{[1 + (\alpha\Psi)^n]^m}, \quad (2.1)$$

$$n = \frac{1}{1 - m}, \quad (2.2)$$

$$\alpha = \frac{1}{h_b}(2^{1/m} - 1)^{1-m}, \quad (2.3)$$

where θ is the volumetric water content, θ_s is the volumetric water content at $s = s_s$, θ_r is the irreducible minimum water content at $s = s_r$, Ψ is the matric potential, m is an experimental parameter based on the soil type and h_b is the bubbling pressure (defined below). For a typical soil-water system, $m \approx 0.5$, $\theta_r \approx 0.1$, $\theta_s \approx 0.5$ and $h_b \approx -355\text{cm}$. Thus, $n \approx 2.0$ and $\alpha \approx 0.005$. A typical curve of water content vs. matric potential for these parameters is given in Figure 2.1.

The bubbling pressure can be defined as follows. At atmospheric pressure, the medium is saturated with $\theta = \theta_s$, where θ_s is the highest value θ can take. Now consider decreasing the matric potential. The medium will remain saturated as the matric potential is decreased until the potential is negative enough that the water will begin to drain. The potential value at which this drainage starts to occur is the bubbling pressure.

For saturated flow, the driving force is again a pressure potential. However, the pressure is now positive and the potential $\Psi > 0$.

If water is allowed to be slightly compressible, the density is not constant and is related to the water pressure through an equation of state, for example,

$$\rho = \rho_0 e^{\beta(p-p_0)}, \quad (2.4)$$

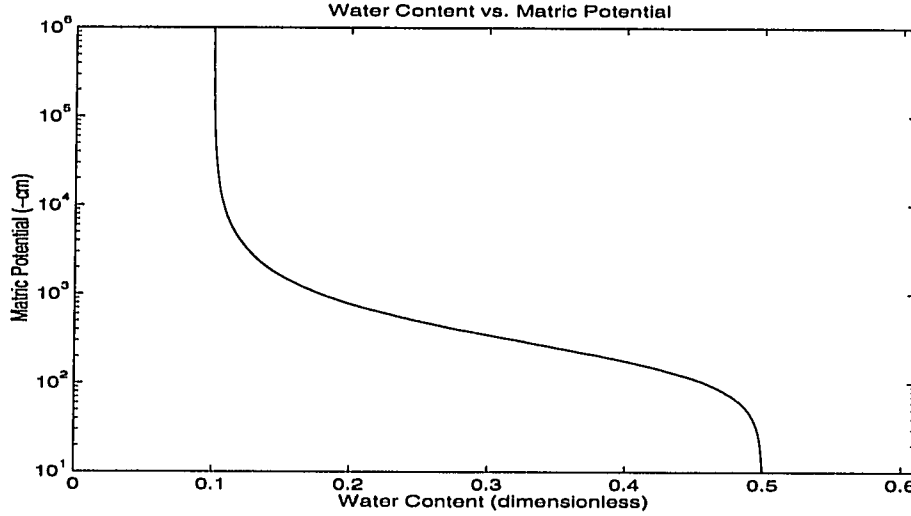


Figure 2.1 Typical van Genuchten curve of water content vs. matric potential.

where ρ_0 is water density at atmospheric pressure p_0 , and β is a small constant. The water compressibility constant, β , is defined as the negative of change in water volume per unit volume per change in pressure, or,

$$\beta = -\frac{dV_w/V_w}{dp} \approx 4.4 \times 10^{-10} \text{m}^2/\text{N}. \quad (2.5)$$

The total soil-moisture potential, Υ , is the sum of the matric potential and a gravity potential. The gravity potential can be expressed as the product of the water density, ρ , the acceleration of gravity, g , and the height, z , above some reference level. Thus, the total soil-moisture potential is,

$$\Upsilon_p = \Psi + \rho g z.$$

If this equation is divided by ρg , the result is the soil moisture potential expressed as energy per unit weight, commonly measured in cm. This potential is,

$$\begin{aligned} \Upsilon_h &= \frac{\Psi}{\rho g} + z \\ &= h + z, \end{aligned}$$

where h is the matric potential expressed in units of length. The matric potential expressed as a length is often referred to as pressure head, and the soil-moisture potential expressed as a length is referred to as hydraulic head.

Darcy's Law for saturated flow and the Buckingham flux law for unsaturated flow relate the flow of water to the gradient of the hydraulic head through the relation,

$$\mathbf{q} = -K(h)\nabla\Upsilon_h, \quad (2.6)$$

where \mathbf{q} is the soil moisture flux (cm/s) and $K(h)$ is the hydraulic conductivity of the soil. The hydraulic conductivity (cm/s) measures the ability of the soil to transmit water. For a saturated medium, the pore space is filled with water and all the pores participate in the transmission of water. Thus, the hydraulic conductivity is a function of position only. However, for an unsaturated medium, some of the pore space is filled with air. Water will only travel through wetted areas, so for an unsaturated medium, the hydraulic conductivity is a function of the moisture content as well as position. Experiments conducted with ideal, uniform porous media have indicated that the hydraulic conductivity can be written as,

$$K(\theta) = \frac{k k_{rw}(\theta) \rho g}{\mu},$$

where k is the intrinsic permeability of the medium (measured in Darcy's where 1 darcy = 10^{-8}cm^2), $k_{rw}(\theta)$ is the relative permeability of water to air (dimensionless) and μ is the dynamic viscosity of water ($\text{N} \cdot \text{s}/\text{cm}^2$). The relative permeability is the ratio of the unsaturated hydraulic conductivity evaluated at θ to the saturated hydraulic conductivity, evaluated at θ_s . The value of k_{rw} is simply a number between 0 and 1. The hydraulic conductivity can also be expressed as a function of the matric potential.

Van Genuchten [38] derived expressions relating the hydraulic conductivity to both the water content and the pressure head. The relationship between K and θ is expressed as,

$$K(\theta) = K_s S_e^{1/2} [1' - (1 - S_e^{1/m})^m]^2,$$

where $S_e = (\theta - \theta_r)/(\theta_s - \theta_r)$ is an effective saturation between 0 and 1, K_s is the saturated hydraulic conductivity and m is the van Genuchten soil parameter. For the typical parameters discussed above, this curve is given in Figure 2.2. The relationship

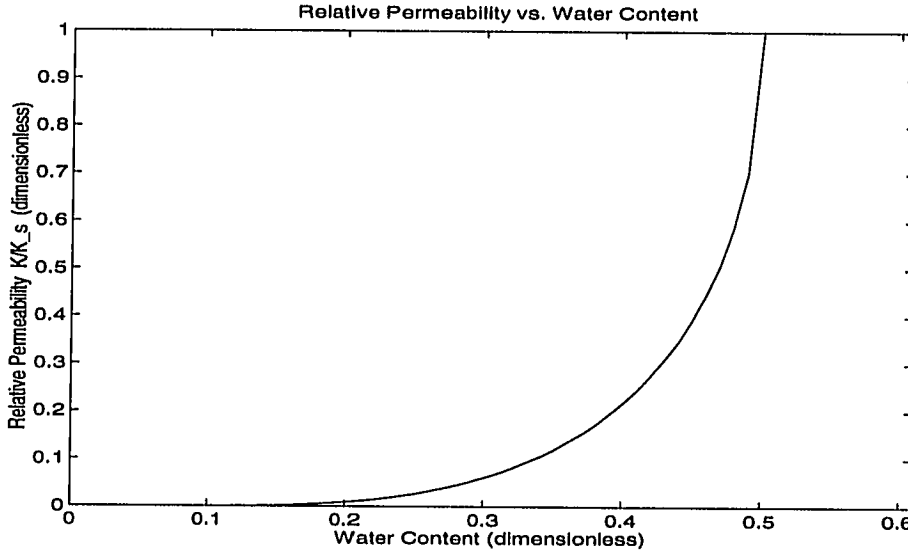


Figure 2.2 Typical van Genuchten curve of hydraulic conductivity vs. water content.

between K and h is,

$$K(h) = K_s \frac{(1 - (\alpha h)^{n-1} [1 + (\alpha h)^n]^{-m})^2}{[1 + (\alpha h)^n]^{m/2}}. \quad (2.7)$$

Figure 2.3 shows this curve for the above described parameters. Note that the relative permeability is just, $K(\theta)/K_s$ or $K(h)/K_s$, so these curves also give the relative permeability function.

Conservation of mass for flow of water in a porous medium requires that the net rate of water mass flow into a small control volume be equal to the time rate of change of water mass storage within the volume plus any source terms. Combining this statement with equation (2.6) gives,

$$\frac{\partial(s\phi\rho)}{\partial t} - \nabla \cdot (\rho K(h) \nabla \Upsilon_h) = f,$$

where f is a source term. This is Richards' equation [55].

The time derivative term can be written as,

$$\frac{\partial(s\phi\rho)}{\partial t} = s\rho \frac{\partial\phi}{\partial t} + s\phi \frac{\partial\rho}{\partial t} + \rho\phi \frac{\partial s}{\partial t}. \quad (2.8)$$

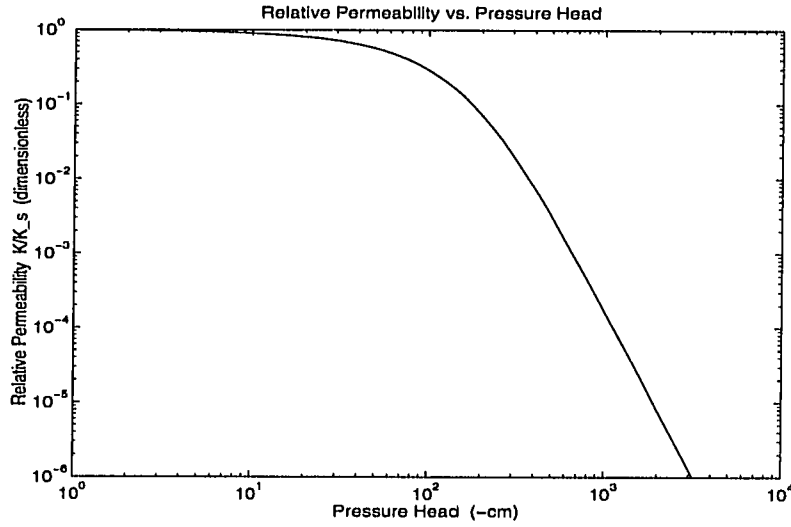


Figure 2.3 Typical van Genuchten curve of hydraulic conductivity vs. pressure head.

Assuming an incompressible medium, ϕ is constant in time and the first term is zero. For unsaturated flow, the second term is small relative to the third and the time derivative is just,

$$\frac{\partial(s\phi\rho)}{\partial t} \approx \rho\phi \frac{\partial s}{\partial t} = \rho \frac{\partial \theta(h)}{\partial t}.$$

For saturated flow, s is a constant equal to s_s , and the second term in (2.8) is the only applicable term. Thus,

$$\frac{\partial(s\phi\rho)}{\partial t} = s\phi \frac{\partial \rho}{\partial t} = s_s\phi\rho\beta \frac{\partial p}{\partial t} = \rho S_s \frac{\partial h}{\partial t},$$

where β is the water compressibility constant in (2.4), and S_s is the specific storage of the aquifer. The specific storage is defined as the volume of water that a unit of aquifer releases from storage under a unit decline in the hydraulic head. In the present work, the component of S_s related to the compressibility of the medium will be ignored. The derivation of S_s is as follows. By equation (2.5),

$$dV_w = -\beta V_w dp.$$

The water volume is just $\phi s_s V_T$ where V_T is the total volume. Assuming $V_T = 1$ and using the relation $p = \rho g h$,

$$dV_w = -\beta \phi s_s \rho g dh.$$

Taking a unit decline in h , $dh = -1$, gives,

$$dV_w = \beta \phi s_s \rho g = S_s.$$

For the spatial derivatives, we have $\nabla \cdot (\rho \mathbf{q})$. This term can be written as,

$$\nabla \cdot (\rho \mathbf{q}) = \nabla \rho \cdot \mathbf{q} + \rho \nabla \cdot \mathbf{q}.$$

Since we assume that water is slightly compressible as in (2.4), $\nabla \rho$ is very small and the first term may be neglected. Canceling the density, we can write Richards' equation as,

$$\frac{\partial \theta(\Upsilon_h)}{\partial t} + S_s \frac{\partial \Upsilon_h}{\partial t} - \nabla \cdot (K(h) \nabla \Upsilon_h) = f, \text{ in } \Omega, \quad (2.9)$$

where the w subscript has been dropped and Ω is the flow domain. For purposes of analysis, θ and K are considered functions of $\Upsilon_h = h + z$. Boundary conditions can be stated as,

$$\Upsilon_h = \Upsilon_D, \text{ on } \Gamma^D, \quad (2.10)$$

$$-K(\Upsilon_h) \nabla \Upsilon_h \cdot \mathbf{n} = g_N, \text{ on } \Gamma^N, \quad (2.11)$$

where $\Gamma^D \cup \Gamma^N = \partial\Omega$, $\Gamma^D \neq \emptyset$, and \mathbf{n} is an outward pointing, unit, normal vector to Ω . This is the mixed form of Richards' equation. The second term in (2.9) is neglected for unsaturated flow, and the first term does not apply for saturated flow. Note here that due to the constant (or passive) air phase pressure assumption, Richards' equation ignores the air phase except through its effects on the hydraulic conductivity, K . An initial condition,

$$\Upsilon_h = \Upsilon^0(x), \quad t = 0, \quad (2.12)$$

completes the specification of the problem.

Owing to the fact that θ is a function of h , we can write equation (2.9) as,

$$(S_s + C(\Upsilon_h)) \frac{\partial \Upsilon_h}{\partial t} - \nabla \cdot (K(\Upsilon_h) \nabla \Upsilon_h) = f, \text{ in } \Omega, \quad (2.13)$$

where $C(\Upsilon_h) = \partial\theta/\partial\Upsilon_h$ denotes the specific moisture capacity. This form is the head-based form of the equation. Lastly, equation (2.9) may be written as,

$$\frac{\partial\theta}{\partial t} - \nabla \cdot (D(\theta)\nabla\theta) = f, \quad \text{in } \Omega, \quad (2.14)$$

where $D(\theta) = K(\theta)/C(\theta)$ is the soil-moisture diffusivity. This is the θ -based form of the equation.

We have now presented a complete mathematical model of partially saturated subsurface flow for a single water phase. In the remaining chapters, we will analyze and solve this model.

Chapter 3

Discretization

In this chapter we present a discussion of spatial and temporal discretization techniques employed in this work. We begin by introducing notation, then presenting variational formulations of Richards' equation. Formulations corresponding to both the mixed and expanded mixed finite element methods will be presented. Discrete approximating spaces and approximation schemes will be discussed. Lastly, comments are made on the time discretization method used.

3.1 Notation

Let Ω be a domain in \mathbb{R}^d with boundary $\Gamma = \partial\Omega$, and let Γ^D be the portion of the boundary where Dirichlet conditions are specified and Γ^N the portion where Neumann conditions are specified. We assume that $\Gamma = \Gamma^D \cup \Gamma^N$. Let $L^2(\Omega)$ be the set of square integrable functions on Ω , i.e., $L^2(\Omega) = \{w \mid \int_{\Omega} w^2 d\Omega < \infty\}$. Let $(L^2(\Omega))^d$ denote the space of d -dimensional vectors which have all components in $L^2(\Omega)$. Furthermore, let $(.,.)$ denote the $L^2(\Omega)$ inner product, scalar and vector, i.e. for $f, g \in L^2(\Omega)$,

$$(f, g) = \int_{\Omega} f \cdot g \, d\Omega.$$

Let $(.,.)_{\partial\Omega}$ denote the $L^2(\partial\Omega)$ inner product and $\|\cdot\|_{\partial\Omega}$ its associated norm.

Let $H(\Omega, \text{div})$ be the space of vectors in $(L^2(\Omega))^d$ which have divergence in $L^2(\Omega)$, i.e. $H(\Omega, \text{div}) = \{\mathbf{v} \in (L^2(\Omega))^d : \nabla \cdot \mathbf{v} \in L^2(\Omega)\}$. If $f \in H(\Omega, \text{div})$, then,

$$\|f\|_{H(\Omega, \text{div})} \equiv \left(\|f\|_{L^2(\Omega)}^2 + \|\nabla \cdot f\|_{L^2(\Omega)}^2 \right)^{1/2}.$$

Let $W_p^k(\Omega)$ be the standard Sobolev space [12, p. 27],

$$W_p^k(\Omega) = \{f : \|f\|_{W_p^k(\Omega)} < \infty\},$$

where,

$$\|f\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha f\|_{L^p(\Omega)}^p \right)^{1/p}.$$

Let $H_s(\Omega)$ for s a positive integer be the Sobolev space, $W_2^s(\Omega)$. Denote the inner product for the H_s Sobolev space as,

$$(f, g)_s = \sum_{|\alpha| \leq s} \int_{\Omega} D^{\alpha} f \cdot D^{\alpha} g \, d\Omega,$$

where $f, g \in H_s(\Omega)$. Let H_{-s} be the dual space of H_s with norm,

$$\|g\|_{-s} = \sup_{\{\Psi \in H_s; \Psi \neq 0\}} \frac{\langle g, \Psi \rangle}{\|\Psi\|_s}, \quad (3.1)$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing between H_s and H_{-s} . We will make use of the fractional Sobolev space $H^{1/2}(\partial\Omega)$ with norm [17],

$$\|g\|_{1/2, \partial\Omega} = \inf_{\{v \in H^1(\Omega); v|_{\partial\Omega} = g\}} \|v\|_{H^1(\Omega)}.$$

Let $\mathbf{V} = H(\Omega, \text{div})$, $\tilde{\mathbf{V}} = (L^2(\Omega))^d$, $W = L^2(\Omega)$ and $\Lambda = H^{1/2}(\partial\Omega)$. Let \mathbf{V}^N and \mathbf{V}^0 denote subspaces of \mathbf{V} with functions whose normal traces on Γ^N are equal to g_N from equation (2.11) and 0, respectively.

3.2 Variational Formulations

To avoid confusion, in this and all following chapters, p will denote hydraulic head. Introducing a velocity variable $\mathbf{u}_M = -K(p)\nabla p$ and writing equation (2.9) as a system of first order equations gives,

$$\frac{\partial \theta(p)}{\partial t} + S_s \frac{\partial p}{\partial t} + \nabla \cdot \mathbf{u}_M = f, \quad (3.2)$$

$$\mathbf{u}_M = -K(p)\nabla p, \quad (3.3)$$

$$p = p_D, \quad \Gamma^D, \quad (3.4)$$

$$\mathbf{u}_M \cdot \mathbf{n} = g_N, \quad \Gamma^N, \quad (3.5)$$

where \mathbf{n} is an outward pointing, unit, normal vector. Multiplying (3.2) by $w \in W$ then integrating and multiplying (3.3) by $K(p)^{-1}$ and $\mathbf{v} \in \mathbf{V}^0$, then integrating by parts, the problem is formulated as finding $(p_M, \mathbf{u}_M) \in (W, \mathbf{V}^N)$ such that,

$$\left(\frac{\partial \theta(p_M)}{\partial t}, w \right) + \left(S_s \frac{\partial p_M}{\partial t}, w \right) + (\nabla \cdot \mathbf{u}_M, w) = (f, w), \quad (3.6)$$

$$(K(p_M)^{-1} \mathbf{u}_M, \mathbf{v}) - (p_M, \nabla \cdot \mathbf{v}) = -(p_D, \mathbf{v} \cdot \mathbf{n})_{\Gamma^D}. \quad (3.7)$$

Equations (3.6)-(3.7) define the variational formulation of the mixed form of Richards' equation corresponding to the mixed finite element method.

For the expanded mixed finite element method we consider a different formulation of the problem,

$$\frac{\partial \theta(p)}{\partial t} + S_s \frac{\partial p}{\partial t} + \nabla \cdot \mathbf{u} = f, \quad (3.8)$$

$$\tilde{\mathbf{u}} = -\nabla p, \quad (3.9)$$

$$\mathbf{u} = K(p)\tilde{\mathbf{u}}, \quad (3.10)$$

$$p = p_D, \Gamma^D, \quad (3.11)$$

$$\mathbf{u} \cdot \mathbf{n} = g_N, \Gamma^N, \quad (3.12)$$

where two additional unknowns, $\tilde{\mathbf{u}}$ and \mathbf{u} have been introduced. Multiplying (3.8) by $w \in W$, multiplying (3.9) by $\mathbf{v} \in \mathbf{V}^0$ and multiplying (3.10) by $\mathbf{v} \in \tilde{\mathbf{V}}$, then integrating each of the resulting equations, the problem can be formulated as finding $(p, \tilde{\mathbf{u}}, \mathbf{u}) \in (W, \tilde{\mathbf{V}}, \mathbf{V}^N)$ such that,

$$\left(\frac{\partial \theta(p)}{\partial t}, w \right) + \left(S_s \frac{\partial p}{\partial t}, w \right) + (\nabla \cdot \mathbf{u}, w) = (f, w), \quad (3.13)$$

$$(\tilde{\mathbf{u}}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) + (p_D, \mathbf{v} \cdot \mathbf{n})_{\Gamma^D} = 0, \quad (3.14)$$

$$(\mathbf{u}, \mathbf{v}) = (K(p)\tilde{\mathbf{u}}, \mathbf{v}). \quad (3.15)$$

Thus, for the expanded mixed method, a set of three equations in three unknowns is solved.

3.3 Approximating Spaces

For mixed finite element methods, the scalar variable p and its velocity are simultaneously approximated. Thus, two approximating spaces are necessary, one for hydraulic head unknowns and one for velocity unknowns. Ideally, these spaces are chosen so that the resulting method has a unique solution.

The approximating spaces used in this work are the Raviart-Thomas-Nedelec spaces on rectangles and parallelepipeds, which are now briefly described. Use of these spaces for linear elliptic problems guarantees a unique solution to the mixed method system [13, 45, 54].

Let \mathcal{T}_h denote a quasi-uniform triangulation of Ω into rectangles with diameter $O(h)$ in two dimensions or parallelepipeds also with diameter $O(h)$ in three dimensions.

The Raviart-Thomas-Nedelec (RTN) [54, 50] approximating space of order k on a rectangular element $E \in \mathcal{T}$ is,

$$\begin{aligned} V_k(E) &= P_{k+1,k}(E) \times P_{k,k+1}(E), \quad d = 2, \\ V_k(E) &= P_{k+1,k,k}(E) \times P_{k,k+1,k} \times P_{k,k,k+1}(E), \quad d = 3, \end{aligned}$$

where $P_{r,s,t}$ is the space of polynomials of degree r in the x direction, s in the y direction and t in the z direction. Raviart and Thomas developed these spaces for two dimensions, and Nedelec for three-dimensions. The space $L^2(\Omega)$ is approximated by,

$$W_k(E) = P_k(E).$$

For the finite difference scheme presented in Chapter 5, we consider the lowest order RTN space, i.e. $k = 0$, on parallelepipeds,

$$\begin{aligned} V_h(E) &= \{(\alpha_1 x_1 + \beta_1, \alpha_2 x_2 + \beta_2, \alpha_3 x_3 + \beta_3)^T : \alpha_i, \beta_i \in \mathbb{R}\}, \\ W_h(E) &= \{\alpha : \alpha \in \mathbb{R}\}, \end{aligned}$$

where the last component in V_h should be deleted in two dimensions. We also define a hybrid space, $\Lambda_h^B \subset L^2(\partial\Omega)$, of Lagrange multipliers for the pressure restricted to $\partial\Omega$ and corresponding to the above RTN spaces [9, 17]. So, on an edge or face e ,

$$\Lambda_h^B(e) = \{\alpha : \alpha \in \mathbb{R}\}.$$

The standard nodal basis is used, where for V_h and Λ_h the nodes are at the midpoints of edges or faces of the elements, and for W_h the nodes are at the centers of the elements. Denote the grid points by

$$(x_{i+1/2}, y_{j+1/2}), i = 0, \dots, N_x, j = 0, \dots, N_y,$$

and define

$$\begin{aligned} x_i &= \frac{1}{2}(x_{i+1/2} + x_{i-1/2}), \quad i = 1, \dots, N_x, \\ y_j &= \frac{1}{2}(y_{j+1/2} + y_{j-1/2}), \quad j = 1, \dots, N_y, \\ h_{i+1/2}^x &= x_{i+1} - x_i, \quad i = 1, \dots, N_x - 1, \\ h_{j+1/2}^y &= y_{j+1} - y_j, \quad j = 1, \dots, N_y - 1, \\ h_i^x &= x_{i+1/2} - x_{i-1/2}, \quad i = 1, \dots, N_x, \\ h_j^y &= y_{j+1/2} - y_{j-1/2}, \quad j = 1, \dots, N_y, \\ h &= \max_{i,j}(h_i^x, h_j^y), \end{aligned}$$

with corresponding notation for a third dimension.

Define discrete inner products corresponding to applications of the midpoint (M), trapezoidal (T) and midpoint by trapezoidal (TM) quadrature rules by

$$\begin{aligned}
(r, s)_M &= \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y r_{ij} s_{ij}, \\
(\mathbf{v}, \mathbf{q})_{TM} &= \sum_{i=0}^{N_x} \sum_{j=1}^{N_y} h_{i+1/2}^x h_j^y v_{i+1/2, j}^x q_{i+1/2, j}^x + \sum_{i=1}^{N_x} \sum_{j=0}^{N_y} h_i^x h_{j+1/2}^y v_{i, j+1/2}^y q_{i, j+1/2}^y, \\
(\mathbf{v}, \mathbf{q})_T &= \sum_{i=0}^{N_x} \sum_{j=1}^{N_y} h_{i+1/2}^x h_j^y \frac{1}{2} (v_{i+1/2, j-1/2}^x q_{i+1/2, j-1/2}^x + v_{i+1/2, j+1/2}^x q_{i+1/2, j+1/2}^x) \\
&\quad + \sum_{i=1}^{N_x} \sum_{j=0}^{N_y} h_i^x h_{j+1/2}^y \frac{1}{2} (v_{i-1/2, j+1/2}^y q_{i-1/2, j+1/2}^y + v_{i+1/2, j+1/2}^y q_{i+1/2, j+1/2}^y),
\end{aligned}$$

where a third sum in each is added for the case of three dimensions. We denote the associated norms by $\|\cdot\|_R$, where $R = M, T$ or TM and by $E_R(\mathbf{q}, \mathbf{r})$, the error in approximating an integral by the given rule, i.e. $E_T(\mathbf{q}, \mathbf{r}) = (\mathbf{q}, \mathbf{r}) - (\mathbf{q}, \mathbf{r})_T$. The error in approximating an integral by either the trapezoidal or the trapezoidal by midpoint rule is [20],

$$|E_Q(\mathbf{q}, \mathbf{v})| \leq C \sum_{E \in \mathcal{T}_h} \sum_{|\alpha|=2} \left\| \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} (\mathbf{q} \cdot \mathbf{v}) \right\|_{L^1(E)} h^2. \quad (3.16)$$

For any $\phi \in L^2(\Omega)$ let $\hat{\phi}$ denote the L^2 projection of ϕ onto W_k , i.e.

$$(\phi, w) = (\hat{\phi}, w), \quad \forall w \in W_k. \quad (3.17)$$

In a similar manner, define an $L^2(\Gamma)$ projection onto Λ_k . These two L^2 projection operators have the following approximation properties for $\phi \in H^{k+1}(\Omega)$ and $\psi \in H^{k+1}(\Gamma)$,

$$\|\hat{\phi} - \phi\| \leq C \|\phi\|_r h^r, \quad 0 \leq r \leq k+1, \quad (3.18)$$

$$\|\hat{\psi} - \psi\|_\Gamma \leq C \|\psi\|_{r, \Gamma} h^r, \quad 0 \leq r \leq k+1, \quad (3.19)$$

$$\|\hat{\psi} - \psi\|_{\Gamma, M} \leq C h^2. \quad (3.20)$$

Associated with the RTN mixed finite element spaces is the projection operator $\Pi : (H^1(\Omega))^d \rightarrow \mathbf{V}_h$, such that for $\mathbf{q} \in H^{k+1}(\Omega)$,

$$(\nabla \cdot \Pi \mathbf{q}, w) = (\nabla \cdot \mathbf{q}, w), \quad \forall w \in W_k, \quad (3.21)$$

$$(\Pi \mathbf{q} \cdot \mathbf{n}, \mu)_e = (\mathbf{q} \cdot \mathbf{n}, \mu)_e, \quad \forall \mu \in \Lambda_k, \quad (3.22)$$

where \mathbf{n} is an outward pointing, unit, normal vector. The following approximation properties hold for the Π projection,

$$\|\mathbf{q} - \Pi\mathbf{q}\| \leq C\|\mathbf{q}\|_r h^r, \quad 0 \leq r \leq k+1, \quad (3.23)$$

$$\|\nabla \cdot (\mathbf{q} - \Pi\mathbf{q})\| \leq C\|\nabla \cdot \mathbf{q}\|_r h^r, \quad 0 \leq r \leq k+1, \quad (3.24)$$

where e is any element edge or face. Note that $\Pi\mathbf{q} \cdot \mathbf{n} = \hat{\mathbf{q}} \cdot \mathbf{n}$ on any boundary edge e .

We will use the following estimate [25] which is true on rectangular or parallelepiped elements. For \mathbf{u} and $\tilde{\mathbf{u}}$ defined by equations (3.13)-(3.15),

$$\|\Pi\mathbf{u} - \mathbf{u}\|_{\text{TM}} + \|\Pi\tilde{\mathbf{u}} - \tilde{\mathbf{u}}\|_{\text{TM}} \leq Ch^2. \quad (3.25)$$

The following inverse estimate for discrete polynomial spaces [12, p. 111] will be used extensively in the following analysis,

Theorem 3.1 Let \mathcal{T}_h be a quasi-uniform triangulation of Ω . Let \hat{E} be a reference element. Let \mathcal{P} be a space of approximating polynomials and $\Upsilon_h = \{y : y|_E \in \mathcal{P}_E \forall E \in \mathcal{T}_h\}$. Then if $\mathcal{P}(\hat{E}) \subseteq W_p^l(\hat{E}) \cap W_q^m(\hat{E})$ where $1 \leq p \leq \infty, 1 \leq q \leq \infty$ and $0 \leq m \leq l$, then there exists a constant Q such that,

$$\left(\sum_{E \in \mathcal{T}_h} \|y\|_{W_p^l(E)}^p \right)^{1/p} \leq Ch^{m-1+\min(0, \frac{d}{p}-\frac{d}{q})} \left(\sum_{E \in \mathcal{T}_h} \|y\|_{W_q^m(E)}^q \right)^{1/q},$$

for all $y \in V_h$. For $s = \infty$, interpret the expression, $\left(\sum_{E \in \mathcal{T}_h} \|y\|_{W_s^m(E)}^s \right)^{1/s}$ as $\max_{E \in \mathcal{T}_h} \|y\|_{W_\infty^l(E)}$.

Let W_h and \mathbf{V}_h be discrete subspaces of W and \mathbf{V} , respectively. Define $V_h^0 = V^0 \cap V_h$ and $V_h^N = V^N \cap V_h$. Then, the continuous time mixed finite element method is to find $(P_M, \mathbf{U}_M) \in (W_h, \mathbf{V}_h^N)$ satisfying,

$$\left(\frac{\partial \theta(P_M)}{\partial t}, w \right) + \left(S_s \frac{\partial P_M}{\partial t}, w \right) + (\nabla \cdot \mathbf{U}_M, w) = (f, w), \quad w \in W_h, \quad (3.26)$$

$$(K(P_M)^{-1} \mathbf{U}_M, \mathbf{v}) - (P_M, \nabla \cdot \mathbf{v}) + (p_D, \mathbf{v} \cdot \mathbf{n})_{\Gamma^D} = 0, \quad \mathbf{v} \in \mathbf{V}_h^0. \quad (3.27)$$

Choosing w in equation (3.26) to be the basis function associated with cell i, j, k , i.e.,

$$w = \begin{cases} 1, & \text{in cell } i, j, k, \\ 0, & \text{otherwise,} \end{cases} \quad (3.28)$$

then equation (3.26) implies that mass is conserved within cell i, j, k . For this reason, the mixed method is known to conserve mass on a cell-by-cell basis.

For the expanded mixed method, we approximate the scalar variable, its velocity and its flux. Thus, three discrete spaces are needed. Let W_h, \mathbf{V}_h and $\tilde{\mathbf{V}}_h$ be discrete subspaces of W, \mathbf{V} and $\tilde{\mathbf{V}}$, respectively. Then the expanded mixed method is formulated as finding $(P, \tilde{\mathbf{U}}, \mathbf{U}) \in (W_h, \tilde{\mathbf{V}}_h, \mathbf{V}_h^N)$ which satisfy,

$$\left(\frac{\partial \theta(P)}{\partial t}, w\right) + \left(S_s \frac{\partial P}{\partial t}, w\right) + (\nabla \cdot \mathbf{U}, w) = (f, w), \quad w \in W_h, \quad (3.29)$$

$$(\tilde{\mathbf{U}}, \mathbf{v}) - (P, \nabla \cdot \mathbf{v}) + (p_D, \mathbf{v} \cdot \mathbf{n})_{\Gamma^D} = 0, \quad \mathbf{v} \in \mathbf{V}_h^0, \quad (3.30)$$

$$(\mathbf{U}, \mathbf{v}) = (K(P)\tilde{\mathbf{U}}, \mathbf{v}), \quad \mathbf{v} \in \tilde{\mathbf{V}}_h. \quad (3.31)$$

For this method, we have the freedom to choose $\tilde{\mathbf{V}}$ not equal to \mathbf{V} . However, for this work $\tilde{\mathbf{V}} = \mathbf{V}$. Note that with w chosen as in (3.28), equation (3.29) again implies conservation of mass over each cell.

3.4 Time Discretization

We consider finding the solution to equation (2.9) over a time interval $J = (0, T)$, where $T > 0$ is some final time. Let $0 = t^0 < t^1 < \dots < t^N = T$ be a given sequence of time steps, $\Delta t^n = t^n - t^{n-1}$ and $\Delta t = \max_n \Delta t^n$. Further, assume that there exist constants c and C such that,

$$c\Delta t^n \leq \Delta t^{n+1} \leq C\Delta t^n, \quad (3.32)$$

for all n .

For $\phi = \phi(t, \cdot)$, let $\phi^n = \phi(t^n, \cdot)$ and denote the discrete and continuous partial derivatives by,

$$d_t \phi^n = \frac{\phi^n - \phi^{n-1}}{\Delta t^n},$$

$$\partial_t \phi = \frac{\partial \phi}{\partial t}.$$

In subsequent chapters, the following norms will be used,

$$\|\phi\|_{L^\infty(J; L^2(\Omega))} \equiv \text{ess sup}_t \|\phi\|(t),$$

$$\|\phi\|_{l^\infty(J; L^2(\Omega))} \equiv \max_{1 \leq n \leq N} \|\phi^n\|,$$

$$\|\phi\|_{L^2(J;L^2(\Omega))} \equiv \left(\int_0^T \|\phi(\cdot, t)\|^2 dt \right)^{\frac{1}{2}},$$

$$\|\phi\|_{l^2(J;L^2(\Omega))} \equiv \left(\sum_{n=1}^N \Delta t^n \|\phi^n\|^2 \right)^{\frac{1}{2}}.$$

This work will use an implicit backward Euler time discretization in order to formulate a discrete time expanded mixed method for Richards' equation. This scheme is to find for each time step $n, n = 1, \dots, N$, functions $(P^n, \tilde{\mathbf{U}}^n, \mathbf{U}^n) \in (W_h, \mathbf{V}_h, \mathbf{V}_h^N)$ satisfying,

$$(d_t \theta(P^n), w) + (S_s d_t P^n, w) + (\nabla \cdot \mathbf{U}^n, w) = (f^n, w), \quad w \in W_h, \quad (3.33)$$

$$(\tilde{\mathbf{U}}^n, \mathbf{v}) - (P^n, \nabla \cdot \mathbf{v}) + (p_D, \mathbf{v} \cdot \mathbf{n})_{\Gamma_D} = 0, \quad \mathbf{v} \in \mathbf{V}_h^0, \quad (3.34)$$

$$(\mathbf{U}^n, \mathbf{v}) = (K(P^n) \tilde{\mathbf{U}}^n, \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_h. \quad (3.35)$$

An implicit method is used to prevent the need for taking unnecessarily small time steps.

The Discrete Gronwall Inequality [30] will be used in coming chapters. We present it now for the sake of completeness,

Lemma 3.1 Let $g(t), f(t)$ and $h(t)$ be nonnegative functions defined on $[0, T]$, $t = i\Delta t, i = 0, \dots, N-1$ and $g(t)$ nondecreasing. If,

$$f(t) + h(t) \leq g(t) + C\Delta t \sum_{s=0}^{t-\Delta t} f(s), \quad (3.36)$$

then,

$$f(t) + h(t) \leq g(t)e^{CT}. \quad (3.37)$$

In conclusion, this chapter has set some notation, introduced the mixed and expanded mixed methods and defined continuous and discrete time numerical schemes for Richards' equation. The next chapter analyzes these schemes.

Chapter 4

An *a priori* Error Analysis of Richards' Equation

In this chapter, we present an error analysis of the expanded mixed finite element method applied to Richards' equation. For simplicity we analyze the form,

$$\frac{\partial \theta(p)}{\partial t} + S_s \frac{\partial p}{\partial t} - \nabla \cdot K(\theta(p)) \nabla p = f, \quad (4.1)$$

where we have made use of the fact that k_{rw} is a function of water content, θ , and have taken $K(\theta(p)) = \{k(x)k_{rw}(\theta(p))\rho g\}/\mu$. Two degenerate conditions can occur. The first, $K = 0$, implies a 0 relative permeability, a condition occurring in very dry media. The second degeneracy occurs when $\theta' = 0$. This condition happens when the media is fully saturated.

We first discuss the case where $K > 0$ for all p and allow for the possibility that $\partial \theta(p)/\partial t = 0$. This situation corresponds to partially to fully saturated flow. For clarity, a continuous time estimate is presented. Bounds for $\|\theta(p) - \theta(P)\|_{L^\infty(L^2)}$ and $\|\tilde{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(L^2)}$ are shown for the partially saturated case, i.e. $S_s = 0$, and a bound on $\|p - P\|_{L^\infty(L^2)}$ for the saturated case. These proofs closely follow the techniques of Arbogast [4] who derived estimates for the case of Galerkin methods. We have extended his work to account for the expanded mixed finite element method.

Next the case where K is bounded above 0 and the derivative of θ is strictly nonzero is considered. This is the case of strictly unsaturated flow. For this situation, optimal convergence of a discrete time scheme is shown.

The next set of estimates are for the case where $K \geq 0$. This situation corresponds to completely dry to fully saturated flow. The Kirchhoff transformation [6, 37, 58] is used to analyze this case. A bound for the flux only is presented in the case that $\partial \theta/\partial p$ can be 0 and a bound for $\|p - P\|_{H_{-1}}$ is presented for $\partial \theta/\partial p \neq 0$. When $K = 0$, the solution generally does not have enough regularity to prove optimal bounds. However, the results presented here would be optimal if the solution had the necessary smoothness. These estimates follow the techniques of Arbogast, Wheeler and Zhang [6] for degenerate equations. This work extends their work to the case of the expanded mixed method and to Richards' equation.

In the following arguments, C will represent a generic constant independent of mesh and time step sizes, and its value should be assumed different at each instance. The arithmetic-geometric inequality,

$$ab \leq \frac{\delta}{2}a^2 + \frac{1}{2\delta}b^2, \quad a, b, \delta \in \mathbb{R}, \quad \delta > 0, \quad (4.2)$$

will be used throughout the analysis.

4.1 Partially to Fully Saturated Flow

In this section, the expanded mixed finite element method applied to Richards' equation (4.1) is considered. The following assumptions are made:

1. The tensor K is symmetric and positive definite.
2. The tensor K is Lipschitz in θ . Thus, there exists a constant L_K independent of two numbers, θ_1 and θ_2 such that, $\|K(\theta_1) - K(\theta_2)\| \leq L_K \|\theta_1 - \theta_2\|$.
3. The function θ is Lipschitz in p . Thus, there exists a constant L_θ independent of two numbers, p_1 and p_2 such that, $\|\theta(p_1) - \theta(p_2)\| \leq L_\theta \|p_1 - p_2\|$.
4. The function $\theta(p)$ is monotone nondecreasing in p .
5. The specific storage S_s can be 0, i.e. $S_s \geq 0$. Recall that $S_s = 0$ in the case of unsaturated flow and $S_s > 0$ for saturated flow.
6. The derivative $\partial_\theta K$ is bounded above, $|\partial_\theta K| \leq C$.
7. The derivative $\partial_t \theta$ is bounded above, $|\partial_t \theta| \leq C$.
8. The composition $(\partial_p \theta) \circ \theta^{-1}$ is Hölder continuous of order β , $0 < \beta \leq 1$. Thus, for $f, g \in \mathbb{R}$,

$$|\partial_p \theta(f) - \partial_p \theta(g)| \leq C |\theta(f) - \theta(g)|^\beta. \quad (4.3)$$

The following analysis bounds the error in the continuous time expanded mixed method applied to Richards' equation given in equations (3.29)-(3.31).

We start with a lemma, proven by Arbogast in [4].

Lemma 4.1 Assume $\theta(x, p)$ is monotone and nondecreasing in $p \in \mathbb{R}$ for each fixed $x \in \Omega$, uniformly Lipschitz in both x and p , and uniformly bounded from above and below. Then, for v and w in \mathbb{R} ,

$$\begin{aligned} (2 \sup_p |\partial_p \theta|)^{-1} (\theta(v) - \theta(w))^2 &\leq \int_w^v (\theta(\wp) - \theta(w)) d\wp \\ &\leq (\theta(v) - \theta(w))(v - w). \end{aligned} \quad (4.4)$$

The following theorem holds for the convergence of the continuous time scheme.

Theorem 4.1 Let $(P, \tilde{\mathbf{U}}, \mathbf{U}) \in (W_h, \mathbf{V}_h, \mathbf{V}_h^N)$ satisfy equations (3.29)-(3.31). Then, under the assumptions given in 1-8 above and for $\delta = 2(k+1)\beta/(1+\beta)$ with $k+1 > d(1+\beta)/(2(3\beta-1)) > 0$,

$$\begin{aligned} \|\theta(p) - \theta(P)\|_{L^\infty(J; L^2(\Omega))} + S_s \|\hat{p} - P\|_{L^\infty(J; L^2(\Omega))}^2 \\ + \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J; L^2(\Omega))} \leq C h^\delta, \end{aligned} \quad (4.5)$$

$$S_s \|p - P\|_{L^\infty(J; L^2(\Omega))}^2 \leq C(h^\delta + h^{k+1}), \quad (4.6)$$

$$\|\tilde{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J; L^2(\Omega))} \leq C(h^\delta + h^k), \quad (4.7)$$

where k is the order of the approximating space.

Proof The proof will be in two parts. First, a bound for $(\theta(p) - \theta(P), p - P)$ is found. Then a bound for $\|\theta(p) - \theta(P)\|$ in terms of $(\theta(p) - \theta(P), p - P)$ will be derived. These two pieces are put together in a continuation argument which gives the final result.

Applying the definitions of the L^2 and Π projections and subtracting (3.29)-(3.31) from (3.13)-(3.15), gives the error equations,

$$(\partial_t(\theta(p) - \theta(P)), w) + (S_s \partial_t(p - P), w) = -(\nabla \cdot (\Pi \mathbf{u} - \mathbf{U}), w), \quad w \in W_h, \quad (4.8)$$

$$(\hat{\mathbf{u}} - \tilde{\mathbf{U}}, \mathbf{v}) = (\hat{p} - P, \nabla \cdot \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_h^0, \quad (4.9)$$

$$(\mathbf{u} - \mathbf{U}, \mathbf{v}) - (K(\theta(p))\tilde{\mathbf{u}} - K(\theta(P))\tilde{\mathbf{U}}, \mathbf{v}) = 0, \quad \mathbf{v} \in \mathbf{V}_h. \quad (4.10)$$

Rewriting (4.10) results in,

$$\begin{aligned} (\Pi \mathbf{u} - \mathbf{U}, \mathbf{v}) &= (\Pi \mathbf{u} - \mathbf{u}, \mathbf{v}) + (K(\theta(p))(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \mathbf{v}) + (K(\theta(p))(\hat{\mathbf{u}} - \tilde{\mathbf{U}}), \mathbf{v}) \\ &\quad + ((K(\theta(p)) - K(\theta(P)))\hat{\mathbf{u}}, \mathbf{v}) \\ &\quad - ((K(\theta(p)) - K(\theta(P)))\tilde{\mathbf{U}}, \mathbf{v}). \end{aligned} \quad (4.11)$$

Let $\gamma = \hat{p} - P$, $\boldsymbol{\eta} = \hat{\mathbf{u}} - \tilde{\mathbf{U}}$ and $\boldsymbol{\xi} = \Pi \mathbf{u} - \mathbf{U}$. Let Q_1 be a constant whose value will be determined later. Then, in (4.8) let $w = \int_t^{\bar{t}} e^{-Q_1 s} \gamma(., s) ds$,

$$\begin{aligned} & \left(\partial_t(\theta(p) - \theta(P)), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(., s) ds \right) + \left(S_s \partial_t(p - P), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(., s) ds \right) \\ & + \left(\nabla \cdot \boldsymbol{\xi}, \int_t^{\bar{t}} e^{-Q_1 s} \gamma(., s) ds \right) = 0. \end{aligned} \quad (4.12)$$

Multiplying (4.9) by $e^{-Q_1 s}$, integrating from t to $\bar{t} < T$ holding \mathbf{v} fixed, then letting $\mathbf{v} = \boldsymbol{\xi}$ gives,

$$\left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds, \boldsymbol{\xi} \right) = \left(\int_t^{\bar{t}} e^{-Q_1 s} \gamma(., s) ds, \nabla \cdot \boldsymbol{\xi} \right). \quad (4.13)$$

In (4.11), let $\mathbf{v} = \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds$ to get,

$$\begin{aligned} & (\boldsymbol{\xi}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds) \\ & = \left(\Pi \mathbf{u} - \mathbf{u}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) + \left(K(\theta(p))(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) \\ & + \left(K(\theta(p))\boldsymbol{\eta}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) + \left((K(\theta(p)) - K(\theta(P)))\hat{\mathbf{u}}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) \\ & - \left((K(\theta(p)) - K(\theta(P)))\boldsymbol{\eta}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right). \end{aligned} \quad (4.14)$$

Combining (4.12)-(4.14) results in,

$$\begin{aligned} & \left(\partial_t(\theta(p) - \theta(P)), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(., s) ds \right) + \left(S_s \partial_t(p - P), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(., s) ds \right) \\ & + \left(\Pi \mathbf{u} - \mathbf{u}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) + \left(K(\theta(p))(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) \\ & + \left(K(\theta(p))\boldsymbol{\eta}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) + \left((K(\theta(p)) - K(\theta(P)))\hat{\mathbf{u}}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) \\ & - \left((K(\theta(p)) - K(\theta(P)))\boldsymbol{\eta}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(., s) ds \right) = 0. \end{aligned} \quad (4.15)$$

Since K is Lipschitz in θ ,

$$|((K(\theta(p)) - K(\theta(P)))\hat{\mathbf{u}}, \mathbf{v})| \leq C \|\theta(p) - \theta(P)\| \|\hat{\mathbf{u}}\|_{L^\infty} \|\mathbf{v}\|. \quad (4.16)$$

Also, by the product rule,

$$\begin{aligned}
K(\theta(p))\boldsymbol{\eta} \cdot \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds &= -\frac{1}{2} \partial_t [K(\theta(p)) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t}] \\
&\quad + \frac{1}{2} [Q_1 K(\theta(p)) + \partial_\theta K(\theta(p)) \partial_t \theta(p)] \times \\
&\quad \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t}
\end{aligned} \tag{4.17}$$

Thus, equation (4.15) becomes,

$$\begin{aligned}
&\left(\partial_t(\theta(p) - \theta(P)), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(\cdot, s) ds \right) + \left(S_s \partial_t(p - P), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(\cdot, s) ds \right) \\
&\quad - \frac{1}{2} \partial_t \int_\Omega K(\theta(p)) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t} dx \\
&\quad + \frac{1}{2} Q_1 \int_\Omega K(\theta(p)) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t} dx \\
&\leq - \left(\Pi \mathbf{u} - \mathbf{u}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right) - \left(K(\theta(p))(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right) \\
&\quad - \frac{1}{2} \int_\Omega \partial_\theta K(\theta(p)) \partial_t \theta(p) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t} dx \\
&\quad + \left((K(\theta(p)) - K(\theta(P))) \boldsymbol{\eta}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right) \\
&\quad + C \|\theta(p) - \theta(P)\| \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|.
\end{aligned} \tag{4.18}$$

Now consider bounding the first four right-hand side terms. Rewriting the first of these gives,

$$\begin{aligned}
|(\Pi \mathbf{u} - \mathbf{u}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds)| &\leq \|\Pi \mathbf{u} - \mathbf{u}\| \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\| \\
&\leq (\|\Pi \mathbf{u} - \mathbf{u}\| e^{-Q_1 t/2}) \left(\left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\| e^{Q_1 t/2} \right) \\
&\leq \epsilon \|\Pi \mathbf{u} - \mathbf{u}\|^2 e^{-Q_1 t} \\
&\quad + C \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|^2 e^{Q_1 t}
\end{aligned} \tag{4.19}$$

Since K is bounded,

$$|(K(\theta(p))(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds)|$$

$$\begin{aligned}
&\leq C \|\tilde{\mathbf{u}} - \hat{\mathbf{u}}\| \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\| \\
&\leq \epsilon \|\tilde{\mathbf{u}} - \hat{\mathbf{u}}\|^2 e^{-Q_1 t} + C \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|^2 e^{Q_1 t}. \quad (4.20)
\end{aligned}$$

Since $\partial_\theta K$ and $\partial_t \theta$ are assumed to be bounded above,

$$\begin{aligned}
&\left| \frac{1}{2} \int_\Omega \partial_\theta K(\theta(p)) \partial_t \theta(p) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 dx e^{Q_1 t} \right| \\
&\leq C \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|^2 e^{Q_1 t}. \quad (4.21)
\end{aligned}$$

Using Theorem 3.1,

$$\begin{aligned}
&|((K(\theta(p)) - K(\theta(P))) \boldsymbol{\eta}, \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds)| \\
&\leq C \|\boldsymbol{\eta}\|_{L^\infty} \|\theta(p) - \theta(P)\| \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\| \\
&\leq C h^{-d/2} \|\boldsymbol{\eta}\| \|\theta(p) - \theta(P)\| \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\| \\
&\leq \epsilon h^{-d} \|\boldsymbol{\eta}\|^2 \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} + C \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|^2 e^{Q_1 t}. \quad (4.22)
\end{aligned}$$

Combining the above bounds with equation (4.18) results in,

$$\begin{aligned}
&\left(\partial_t(\theta(p) - \theta(P)), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(\cdot, s) ds \right) + \left(S_s \partial_t(p - P), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(\cdot, s) ds \right) \\
&- \frac{1}{2} \partial_t \int_\Omega K(\theta(p)) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t} dx \\
&+ \frac{1}{2} Q_1 \int_\Omega K(\theta(p)) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t} dx \\
&\leq C \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|^2 e^{Q_1 t} + \epsilon \{1 + h^{-d} \|\boldsymbol{\eta}\|^2\} \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} \\
&+ \epsilon \{\|\Pi \mathbf{u} - \mathbf{u}\|^2 + \|\tilde{\mathbf{u}} - \hat{\mathbf{u}}\|^2\} e^{-Q_1 t}. \quad (4.23)
\end{aligned}$$

This equation is integrated in time over $(0, \bar{t})$. The first two left-hand side terms are handled by integration by parts,

$$\begin{aligned}
&\int_0^{\bar{t}} \left(\partial_t(\theta(p) - \theta(P)), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(\cdot, s) ds \right) dt \\
&= - \left(\theta(p^0) - \theta(P^0), \int_0^{\bar{t}} \gamma(\cdot, t) e^{-Q_1 t} dt \right) + \int_0^{\bar{t}} (\theta(p) - \theta(P), p - P) e^{-Q_1 t} dt
\end{aligned}$$

$$\begin{aligned}
& - \int_0^{\bar{t}} (\theta(p) - \theta(P), p - \hat{p}) e^{-Q_1 t} dt \\
& \geq \int_0^{\bar{t}} (\theta(p) - \theta(P), p - P) e^{-Q_1 t} dt - C \left\{ \int_0^{\bar{t}} \|p - \hat{p}\|^2 e^{-Q_1 t} dt + \|\theta(p^0) - \theta(P^0)\|^2 \right\} \\
& \quad - \epsilon \left\{ \int_0^{\bar{t}} \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt + \left\| \int_0^{\bar{t}} \gamma(\cdot, t) e^{-Q_1 t} dt \right\|^2 \right\}. \tag{4.24}
\end{aligned}$$

The second term is,

$$\begin{aligned}
& \int_0^{\bar{t}} S_s (\partial_t (\hat{p} - P), \int_t^{\bar{t}} e^{-Q_1 s} \gamma(\cdot, s) ds) dt \\
& \geq S_s \int_0^{\bar{t}} \|\gamma\|^2 e^{-Q_1 t} dt - C S_s \|\hat{p}^0 - P^0\|^2 - \epsilon S_s \left\| \int_0^{\bar{t}} \gamma(\cdot, s) e^{-Q_1 s} ds \right\|^2. \tag{4.25}
\end{aligned}$$

So, integrating (4.23) over $(0, \bar{t})$ gives,

$$\begin{aligned}
& \int_0^{\bar{t}} (\theta(p) - \theta(P), p - P) e^{-Q_1 t} dt + S_s \int_0^{\bar{t}} \|\hat{p} - P\|^2 e^{-Q_1 t} dt \\
& \quad + \frac{1}{2} \int_{\Omega} K(\theta(p^0)) \left(\int_0^{\bar{t}} \boldsymbol{\eta}(\cdot, s) e^{-Q_1 s} ds \right)^2 dx \\
& \quad + \frac{1}{2} Q_1 \int_0^{\bar{t}} \int_{\Omega} K(\theta(p)) \left(\int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right)^2 e^{Q_1 t} dx dt \\
& \leq C \int_0^{\bar{t}} \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|^2 e^{Q_1 t} dt + C \int_0^{\bar{t}} \|p - \hat{p}\|^2 e^{-Q_1 t} dt + C \|\theta(p^0) - \theta(P^0)\|^2 \\
& \quad + C S_s \|\hat{p}^0 - P^0\|^2 + \epsilon \left\{ \int_0^{\bar{t}} \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt + \left\| \int_0^{\bar{t}} \gamma(\cdot, t) e^{-Q_1 t} dt \right\|^2 \right\} \\
& \quad + \epsilon \int_0^{\bar{t}} \{1 + h^{-d} \|\boldsymbol{\eta}\|^2\} \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt \\
& \quad + \epsilon \int_0^{\bar{t}} \{\|\Pi \mathbf{u} - \mathbf{u}\|^2 + \|\tilde{\mathbf{u}} - \hat{\mathbf{u}}\|^2\} e^{-Q_1 t} dt. \tag{4.26}
\end{aligned}$$

Since θ is Lipschitz in p ,

$$\begin{aligned}
\epsilon \int_0^{\bar{t}} \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt & \leq \epsilon \int_0^{\bar{t}} \|p - P\|^2 e^{-Q_1 t} dt \\
& \leq \epsilon \int_0^{\bar{t}} \|p - \hat{p}\|^2 e^{-Q_1 t} dt \\
& \quad + \epsilon \int_0^{\bar{t}} \|\gamma(\cdot, t)\|^2 e^{-Q_1 t} dt. \tag{4.27}
\end{aligned}$$

Choosing Q_1 to exactly cancel the last left-hand side term in (4.26) with the first right-hand side term, again using the fact that θ is Lipschitz and choosing ϵ small

enough gives,

$$\begin{aligned}
& \int_0^{\bar{t}} (\theta(p) - \theta(P), p - P) e^{-Q_1 t} dt + S_s \int_0^{\bar{t}} \|\hat{p} - P\|^2 e^{-Q_1 t} dt \\
& + \frac{1}{2} \int_{\Omega} K(\theta(p^0)) \left(\int_0^{\bar{t}} \boldsymbol{\eta}(\cdot, s) e^{-Q_1 s} ds \right)^2 dx \\
& \leq C \int_0^{\bar{t}} \|p - \hat{p}\|^2 e^{-Q_1 t} dt + C \|\theta(p^0) - \theta(P^0)\|^2 + C S_s \|\hat{p}^0 - P^0\|^2 \\
& + \epsilon \left\| \int_0^{\bar{t}} \gamma(\cdot, t) e^{-Q_1 t} dt \right\|^2 \\
& + \epsilon h^{-d} \int_0^{\bar{t}} \|\boldsymbol{\eta}\|^2 \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt \\
& + \epsilon \int_0^{\bar{t}} \{\|\Pi \mathbf{u} - \mathbf{u}\|^2 + \|\tilde{\mathbf{u}} - \hat{\mathbf{u}}\|^2\} e^{-Q_1 t} dt \\
& + \epsilon \int_0^{\bar{t}} \|\gamma(\cdot, t)\|^2 e^{-Q_1 t} dt.
\end{aligned} \tag{4.28}$$

Before continuing, we present the following lemma.

Lemma 4.2 For P and $\tilde{\mathbf{U}}$ defined as in equations (3.29)-(3.31) and for $\gamma = \hat{p} - P$ and $\boldsymbol{\eta} = \hat{\mathbf{u}} - \tilde{\mathbf{U}}$, we have,

$$\|\gamma\| \leq C \|\boldsymbol{\eta}\|, \tag{4.29}$$

$$\left\| \int_t^{\bar{t}} e^{-Q_1 s} \gamma(\cdot, s) ds \right\| \leq C \left\| \int_t^{\bar{t}} e^{-Q_1 s} \boldsymbol{\eta}(\cdot, s) ds \right\|. \tag{4.30}$$

Proof Let ϕ satisfy the equation $-\Delta \phi = \gamma$, and let f satisfy,

$$f = -\nabla \phi, \text{ in } \Omega, \tag{4.31}$$

$$f \cdot \mathbf{n} = 0, \text{ on } \Gamma^N, \tag{4.32}$$

$$\phi = 0, \text{ on } \Gamma^D, \tag{4.33}$$

where we have assumed that $\Gamma^D \neq \emptyset$. Then, recalling (4.9),

$$\begin{aligned}
\|\gamma\|^2 &= (\gamma, \nabla \cdot f) = (\gamma, \nabla \cdot \Pi f) = (\boldsymbol{\eta}, \Pi f) \\
&\leq \|\boldsymbol{\eta}\| (\|\Pi f - f\| + \|f\|) \\
&\leq \|\boldsymbol{\eta}\| (Ch \|f\|_1 + \|f\|) \\
&\leq C \|\boldsymbol{\eta}\| \|\phi\|_2 \\
&\leq C \|\boldsymbol{\eta}\| \|\gamma\|,
\end{aligned}$$

where the last inequality holds by elliptic regularity. In a similar manner (4.30) is shown. \square

Applying approximation results, (3.18) and (3.23), and Lemma 4.2 to equation (4.28) and noting that $\int_0^{\bar{t}} e^{-Q_1 t} dt = \frac{1}{Q_1} (1 - \frac{1}{e^{Q_1 \bar{t}}}) = C$, gives,

$$\begin{aligned} & \int_0^{\bar{t}} (\theta(p) - \theta(P), p - P) e^{-Q_1 t} dt + S_s \int_0^{\bar{t}} \|\hat{p} - P\|^2 e^{-Q_1 t} dt \\ & + \frac{1}{2} \int_{\Omega} K(\theta(p^0)) \left(\int_0^{\bar{t}} \boldsymbol{\eta}(\cdot, s) e^{-Q_1 s} ds \right)^2 dx \\ & \leq C h^{2(k+1)} + C \|\theta(p^0) - \theta(P^0)\|^2 + C S_s \|\hat{p}^0 - P^0\|^2 \\ & + \epsilon h^{-d} \int_0^{\bar{t}} \|\boldsymbol{\eta}\|^2 \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt. \end{aligned} \quad (4.34)$$

This completes the first part of the proof. We come back to this estimate later.

Let $w = \gamma$ in (4.8), $\mathbf{v} = \boldsymbol{\eta}$ in (4.9) and $\mathbf{v} = \boldsymbol{\xi}$ in (4.10) and combine the three resulting equations to get,

$$\begin{aligned} & (\partial_t(\theta(p) - \theta(P)), \hat{p} - P) + (S_s \partial_t(\hat{p} - P), \hat{p} - P) + (\Pi \mathbf{u} - \mathbf{u}, \boldsymbol{\eta}) \\ & + (K(\theta(p)) \tilde{\mathbf{u}} - K(\theta(P)) \tilde{\mathbf{U}}, \boldsymbol{\eta}) = 0. \end{aligned} \quad (4.35)$$

Now,

$$\begin{aligned} K(\theta(p)) \tilde{\mathbf{u}} - K(\theta(P)) \tilde{\mathbf{U}} &= K(\theta(p))(\tilde{\mathbf{u}} - \hat{\tilde{\mathbf{u}}}) + K(\theta(P))(\hat{\tilde{\mathbf{u}}} - \tilde{\mathbf{U}}) \\ &+ (K(\theta(p)) - K(\theta(P))) \hat{\tilde{\mathbf{u}}}. \end{aligned} \quad (4.36)$$

Combining equations (4.36) and (4.35), applying approximation properties (3.18) and (3.23), using the assumption that K is Lipschitz and using the arithmetic-geometric inequality,

$$\begin{aligned} & (\partial_t(\theta(p) - \theta(P)), p - P) + (S_s \partial_t(\hat{p} - P), \hat{p} - P) + \frac{1}{2} (K(\theta(P)) \boldsymbol{\eta}, \boldsymbol{\eta}) \\ & = -(\Pi \mathbf{u} - \mathbf{u}, \boldsymbol{\eta}) - (K(\theta(p))(\tilde{\mathbf{u}} - \hat{\tilde{\mathbf{u}}}), \boldsymbol{\eta}) - ((K(\theta(p)) - K(\theta(P))) \hat{\tilde{\mathbf{u}}}, \boldsymbol{\eta}) \\ & \quad + (\partial_t(\theta(p) - \theta(P)), p - \hat{p}) \\ & \leq C h^{2(k+1)} + C \|\theta(p) - \theta(P)\|^2 + (\partial_t(\theta(p) - \theta(P)), p - \hat{p}). \end{aligned} \quad (4.37)$$

Note that,

$$\begin{aligned} \partial_t \int_P^p (\theta(\wp) - \theta(P)) d\wp &= (\theta(p) - \theta(P)) \partial_t p - \partial_t \theta(P) (p - P) \\ &= (\theta(p) - \theta(P)) \partial_t p + \partial_t [\theta(p) - \theta(P)] (p - P) \\ &\quad - \partial_t \theta(p) (p - P). \end{aligned} \quad (4.38)$$

So,

$$\begin{aligned}
& \partial_t[\theta(p) - \theta(P)](p - P)e^{-Q_2 t} \\
&= \partial_t\left[\int_P^p (\theta(\wp) - \theta(P))d\wp e^{-Q_2 t}\right] + Q_2 \int_P^p (\theta(\wp) - \theta(P))d\wp e^{-Q_2 t} \\
&+ \{\partial_t \theta(p)(p - P) - (\theta(p) - \theta(P))\partial_t p\}e^{-Q_2 t}.
\end{aligned} \tag{4.39}$$

Consider the last term in this equation and apply the chain rule and Mean Value Theorem, where for some $w \in (P, p)$,

$$\begin{aligned}
& |\partial_t \theta(p)(p - P) - (\theta(p) - \theta(P))\partial_t p| \\
&= |(\partial_p \theta(p) - \partial_p \theta(w))(p - P)\partial_t p| \\
&\leq C|\theta(p) - \theta(w)|^\beta |p - P| \\
&\leq C|\theta(p) - \theta(P)|^\beta |p - P| \\
&\leq C|(\theta(p) - \theta(P))(p - P)|^{2\beta/(1+\beta)} + \epsilon\{|\hat{p} - P|^2 + |p - \hat{p}|^2\},
\end{aligned} \tag{4.40}$$

where we have used the inequality, [4]

$$|ab| \leq \frac{1}{\epsilon^{p/q} p} |a|^p + \frac{\epsilon}{q} |b|^q, \tag{4.41}$$

for any $1 < p < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$, which implies, $|a|^\beta |b| \leq C|ab|^{2\beta/(1+\beta)} + \epsilon b^2$.

Multiplying (4.37) by $e^{-Q_2 t}$, combining with the above bounds and integrating from 0 to t gives,

$$\begin{aligned}
& \int_0^t \int_\Omega \partial_\tau \left[\int_P^p (\theta(\wp) - \theta(P))d\wp \right] e^{-Q_2 \tau} dx d\tau + Q_2 \int_0^t \int_\Omega \int_P^p (\theta(\wp) - \theta(P))d\wp e^{-Q_2 \tau} dx d\tau \\
&+ \int_0^t (S_s \partial_\tau (\hat{p} - P), \hat{p} - P) e^{-Q_2 \tau} d\tau + \frac{1}{2} \int_0^t (K(\theta(P))\boldsymbol{\eta}, \boldsymbol{\eta}) e^{-Q_2 \tau} d\tau \\
&\leq Ch^{2(k+1)} + C \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau + \int_0^t (\partial_t(\theta(p) - \theta(P)), p - \hat{p}) e^{-Q_2 \tau} d\tau \\
&+ \int_\Omega \int_0^t |(\theta(p) - \theta(P))(p - P)|^{2\beta/(1+\beta)} e^{-Q_2 \tau} d\tau dx \\
&+ \int_\Omega \epsilon \int_0^t (|\hat{p} - P|^2 + |p - \hat{p}|^2) e^{-Q_2 \tau} d\tau dx.
\end{aligned} \tag{4.42}$$

Lemma (4.1) implies that, $C(\theta(v) - \theta(w))^2 \leq \int_w^v (\theta(\mu) - \theta(w))d\mu$. So, the above becomes,

$$C \int_0^t (\partial_\tau \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau + Q_2 \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau$$

$$\begin{aligned}
& + \int_0^t (S_s \partial_\tau (\hat{p} - P), \hat{p} - P) e^{-Q_2 \tau} d\tau + \frac{1}{2} \int_0^t (K(\theta(P)) \boldsymbol{\eta}, \boldsymbol{\eta}) e^{-Q_2 \tau} d\tau \\
& \leq C h^{2(k+1)} + C \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau + \int_0^t (\partial_\tau (\theta(p) - \theta(P)), p - \hat{p}) e^{-Q_2 \tau} d\tau \\
& + \int_0^t |(\theta(p) - \theta(P), p - P)|^{2\beta/(1+\beta)} e^{-Q_2 \tau} d\tau + \epsilon \int_0^t \|\hat{p} - P\|^2 e^{-Q_2 \tau} d\tau \\
& + \epsilon \int_0^t \|p - \hat{p}\|^2 e^{-Q_2 \tau} d\tau
\end{aligned} \tag{4.43}$$

By integration by parts,

$$\begin{aligned}
& \int_0^t (\partial_\tau (\theta(p) - \theta(P)), p - \hat{p}) e^{-Q_2 \tau} d\tau \\
& = - \int_0^t (\theta(p) - \theta(P), \partial_\tau (p - \hat{p})) e^{-Q_2 \tau} d\tau + Q_2 \int_0^t (\theta(p) - \theta(P), p - \hat{p}) e^{-Q_2 \tau} d\tau \\
& + (\theta(p) - \theta(P), p - \hat{p}) e^{-Q_2 t} - (\theta(p^0) - \theta(P^0), p^0 - \hat{p}^0) \\
& \leq \epsilon \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau + C \int_0^t \|\partial_\tau (p - \hat{p})\|^2 e^{-Q_2 \tau} d\tau \\
& + \epsilon Q_2 \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau + Q_2 \int_0^t \|p - \hat{p}\|^2 e^{-Q_2 \tau} d\tau + \hat{\epsilon} \|\theta(p) - \theta(P)\|^2 e^{-Q_2 t} \\
& + C \|p - \hat{p}\|^2 e^{-Q_2 t} + |(\theta(p^0) - \theta(P^0), p^0 - \hat{p}^0)|.
\end{aligned} \tag{4.44}$$

Also by integration by parts,

$$\begin{aligned}
C \int_0^t (\partial_\tau \|\theta(p) - \theta(P)\|^2) e^{-Q_2 \tau} d\tau & = Q_2 C \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau \\
& + C \|\theta(p) - \theta(P)\|^2 e^{-Q_2 t} \\
& - C \|\theta(p^0) - \theta(P^0)\|^2,
\end{aligned} \tag{4.45}$$

and

$$\begin{aligned}
\int_0^t (S_s \partial_\tau (\hat{p} - P), \hat{p} - P) e^{-Q_2 \tau} d\tau & = \int_0^t \frac{S_s}{2} \partial_\tau \|\hat{p} - P\|^2 e^{-Q_2 \tau} d\tau \\
& = \frac{S_s Q_2}{2} \int_0^t \|\hat{p} - P\|^2 e^{-Q_2 \tau} d\tau + \frac{S_s}{2} \|\hat{p} - P\|^2 e^{-Q_2 t} \\
& - \frac{S_s}{2} \|\hat{p}^0 - P^0\|^2.
\end{aligned} \tag{4.46}$$

Combining (4.44) and (4.45) with (4.43) gives,

$$\begin{aligned}
& Q_2 C \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau + C \|\theta(p) - \theta(P)\|^2 e^{-Q_2 t} + \frac{S_s Q_2}{2} \int_0^t \|\hat{p} - P\|^2 e^{-Q_2 \tau} d\tau \\
& + \frac{S_s}{2} \|\hat{p} - P\|^2 e^{-Q_2 t} + \frac{1}{4} \int_0^t (K(\theta(P)) \boldsymbol{\eta}, \boldsymbol{\eta}) e^{-Q_2 \tau} d\tau
\end{aligned}$$

$$\begin{aligned}
&\leq Ch^{2(k+1)} + (\epsilon Q_2 + C) \int_0^t \|\theta(p) - \theta(P)\|^2 e^{-Q_2 \tau} d\tau + C \int_0^t \|\partial_\tau(p - \hat{p})\|^2 e^{-Q_2 \tau} d\tau \\
&\quad + \int_0^t |(\theta(p) - \theta(P), p - P)|^{2\beta/(1+\beta)} e^{-Q_2 \tau} d\tau + \hat{\epsilon} \|\theta(p) - \theta(P)\|^2 e^{-Q_2 t} \\
&\quad + |(\theta(p^0) - \theta(P^0), p^0 - \hat{p}^0)| + C \|\theta(p^0) - \theta(P^0)\|^2 + \frac{S_s}{2} \|\hat{p}^0 - P^0\|^2.
\end{aligned} \tag{4.47}$$

So, choosing Q_2 and ϵ to make the first left-hand side term cancel the second right-hand side term, noting that the third left-hand side term is nonnegative and taking $\hat{\epsilon}$ small enough,

$$\begin{aligned}
&C \|\theta(p) - \theta(P)\|^2 e^{-Q_2 t} + \frac{S_s}{2} \|\hat{p} - P\|^2 e^{-Q_2 \tau} + \frac{1}{4} \int_0^t (K(\theta(P))\boldsymbol{\eta}, \boldsymbol{\eta}) e^{-Q_2 \tau} d\tau \\
&\leq Ch^{2(k+1)} + \int_0^t |(\theta(p) - \theta(P), p - P)|^{2\beta/(1+\beta)} e^{-Q_2 \tau} d\tau + |(\theta(p^0) - \theta(P^0), p^0 - \hat{p}^0)| \\
&\quad + C \|\theta(p^0) - \theta(P^0)\|^2 + \frac{S_s}{2} \|\hat{p}^0 - P^0\|^2.
\end{aligned} \tag{4.48}$$

Take $P^0 = \hat{p}^0$. Then,

$$\|\theta(p^0) - \theta(P^0)\|^2 \leq C(\|p^0 - \hat{p}^0\|^2 + \|\hat{p}^0 - P^0\|^2) \leq Ch^{2(k+1)}.$$

Thus,

$$\begin{aligned}
&C \|\theta(p) - \theta(P)\|^2 e^{-Q_2 t} + \frac{S_s}{2} \|\hat{p} - P\|^2 e^{-Q_2 \tau} + \frac{1}{4} \int_0^t (K(\theta(P))\boldsymbol{\eta}, \boldsymbol{\eta}) e^{-Q_2 \tau} d\tau \\
&\leq C\{h^{2(k+1)} + \left(\int_0^t |(\theta(p) - \theta(P), p - P)| e^{-Q_2(1+\beta)\tau/2\beta} d\tau\right)^{2\beta/(1+\beta)}\},
\end{aligned} \tag{4.49}$$

where we have used the Hölder inequality with $p = (1+\beta)/2\beta$ and $q = (1-\beta)/(1+\beta)$. This completes the second part of the proof. We now combine the first two parts to derive the desired estimate.

For some fixed value C_0 independent of h , let $T' \leq T$ be the largest value of time for which,

$$\|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J'; L^2(\Omega))} \leq C_0 h^{d\beta/(3\beta-1)}, \tag{4.50}$$

where $J' = (0, T')$ and d is the spatial dimension. Since initially $\hat{\mathbf{u}} - \tilde{\mathbf{U}} = 0$ and $\tilde{\mathbf{u}}$ is assumed continuous in time, we must have $T' > 0$.

We make use of the following inequality [4]. If $\frac{1}{2} < \delta \leq 1$, then

$$|abc|^\delta \leq |b| + (|a|^{\delta/(2\delta-1)}|b|)^{(2\delta-1)/\delta}|c|. \tag{4.51}$$

Note that since $0 < \beta \leq 1$, $k + 1 \geq \frac{2(k+1)\beta}{1+\beta}$.

Let $t = \bar{t}$ be the time when $\text{ess sup}_{t \in (0, J')} \|\theta(p) - \theta(P)\|$ is attained. Then, combining equations (4.34) and (4.49),

$$\begin{aligned}
& \|\theta(p) - \theta(P)\|_{L^\infty(J'; L^2(\Omega))}^2 + S_s \|\hat{p} - P\|_{L^\infty(J'; L^2(\Omega))}^2 + \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(\bar{J}; L^2(\Omega))}^2 \\
& \leq Ch^{(k+1)(4\beta/(1+\beta))} + \left(\epsilon h^{-d} \int_0^{\bar{t}} \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|^2 \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt \right)^{2\beta/(1+\beta)} \\
& \leq Ch^{(k+1)(4\beta/(1+\beta))} \\
& \quad + (\epsilon h^{-d} \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(\bar{J}; L^2(\Omega))}^2 \|\theta(p) - \theta(P)\|_{L^\infty(J'; L^2(\Omega))}^2)^{2\beta/(1+\beta)} \\
& \leq Ch^{(k+1)(4\beta/(1+\beta))} + \epsilon^{2\beta/(1+\beta)} \{ \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(\bar{J}; L^2)}^2 \\
& \quad + ((h^{-d2\beta/(3\beta-1)}) \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(\bar{J}; L^2)}^2)^{(3\beta-1)/2\beta} \|\theta(p) - \theta(P)\|_{L^\infty(J'; L^2)}^2 \}, \tag{4.52}
\end{aligned}$$

where $\bar{J} = (0, \bar{t})$, inequality (4.51) is used to derive the last inequality, and we have assumed $3\beta > 1$ so that $\delta = 2\beta/(1+\beta) > \frac{1}{2}$. In (4.50), we took the exponent on h to be exactly large enough to cancel the $-d$ exponent in (4.52). Now, since $\bar{t} \leq t'$, we can use (4.50) and take ϵ small enough to hide the last two right-hand side terms. Thus,

$$\begin{aligned}
& \|\theta(p) - \theta(P)\|_{L^\infty(J'; L^2)}^2 + S_s \|\hat{p} - P\|_{L^\infty(J'; L^2(\Omega))}^2 + \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(\bar{J}; L^2)}^2 \\
& \leq Ch^{(k+1)(4\beta/(1+\beta))}. \tag{4.53}
\end{aligned}$$

Without a bound on $\|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|$ or $\|\theta(p) - \theta(P)\|$, we could not have hidden the last right-hand side term. We thus assumed the minimum for one bound and derived the other. We now can improve the first.

Let $t = T'$ and again combine equations (4.34) and (4.49),

$$\begin{aligned}
& \|\theta(p) - \theta(P)\|_{L^\infty(J'; L^2(\Omega))}^2 + S_s \|\hat{p} - P\|_{L^\infty(J'; L^2(\Omega))}^2 + \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J'; L^2(\Omega))}^2 \\
& \leq Ch^{(k+1)(4\beta/(1+\beta))} + \epsilon^{2\beta/(1+\beta)} \{ \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J'; L^2)}^2 \\
& \quad + ((h^{-2\beta d/(3\beta-1)}) \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J'; L^2)}^2)^{3\beta-1/2\beta} \|\theta(p) - \theta(P)\|_{L^\infty(J'; L^2)}^2 \}. \tag{4.54}
\end{aligned}$$

Using the bounds (4.53) and (4.50) and taking ϵ small enough gives,

$$\|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J'; L^2)}^2 \leq Ch^{(k+1)(4\beta/(1+\beta))}. \tag{4.55}$$

We continue by contradiction. Suppose that $T' < T$, that $h_0 > 0$ is fixed and that $k + 1 > d(1+\beta)/(2(3\beta-1)) > 0$. Then,

$$\|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J'; L^2)} \leq Ch^{2(k+1)\beta/(1+\beta)} \leq \frac{1}{2} C_0 h^{d\beta/(3\beta-1)}, \tag{4.56}$$

for all values of $h < h_0$. Since $T' < T$ and T' is the maximal value such that (4.50) is true, we have a contradiction. Thus, $T' = T$ and,

$$\|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J;L^2)}^2 \leq Ch^{(k+1)(4\beta/(1+\beta))}. \quad (4.57)$$

This, together with (4.53) gives the desired result. \square

We have the following corollary,

Corollary 4.1 For the scheme given by (3.29)-(3.31) and $\bar{\delta} = k+1$ with $k+1 > d(1+\beta)/(2(3\beta-1)) > 0$, we have,

$$\left(\int_0^T (\theta(p) - \theta(P), p - P) dt \right)^{\frac{1}{2}} \leq Ch^{\bar{\delta}}. \quad (4.58)$$

Proof Let $\bar{t} = T$ in (4.34) to get,

$$\begin{aligned} & \int_0^T (\theta(p) - \theta(P), p - P) e^{-Q_1 t} dt \\ & \leq Ch^{2(k+1)} + \epsilon h^{-d} \int_0^T \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|^2 \|\theta(p) - \theta(P)\|^2 e^{-Q_1 t} dt \\ & \leq Ch^{2(k+1)} + \epsilon h^{-d} \|\theta(p) - \theta(P)\|_{L^\infty(J;L^2(\Omega))}^2 \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}\|_{L^2(J;L^2(\Omega))}^2 \\ & \leq Ch^{2(k+1)} + \epsilon h^{2(k+1)4\beta/(1+\beta)-d} \\ & \leq Ch^{2(k+1)} + \epsilon h^{2(k+1)4\beta/(1+\beta)-(k+1)2(3\beta-1)/(1+\beta)} \\ & \leq Ch^{2(k+1)}. \end{aligned} \quad (4.59)$$

\square

As Arbogast [4] points out, the nonlinear form $\left(\int_0^T (\theta(p) - \theta(P), p - P) dt \right)^{\frac{1}{2}}$ tells us something about the error of the scheme, since for two constants c and C ,

$$c|\theta(p) - \theta(P)| \leq ((\theta(p) - \theta(P))(p - P))^{\frac{1}{2}} \leq C|p - P|.$$

Thus, as the nonlinear form gets smaller, the error in the water content also decreases. Furthermore, the bound on the nonlinear form is optimal, since $O(h^{k+1})$ is the order of truncation error for approximation with polynomials of order k .

The estimates in this section show bounds for the case where the equation is degenerate. For this case, S_s can be 0, and we have only a bound on $\|\theta(p) - \theta(P)\|$. In the next section we make a simplifying assumption that allows us to bound the error in the hydraulic head directly.

4.2 Strictly Partially Saturated Flow

In this section, the case of strictly partially saturated flow is analyzed. It is assumed that the flow never reaches the fully saturated realm and thus that the derivative, $\partial\theta/\partial p$ is never zero.

The following assumptions are made.

1. The tensor K is symmetric, positive definite and bounded, i.e. $K_* \leq K_{ij} \leq K^*$ for all i and j , where K_* and K^* are fixed positive constants.
2. The tensor K is Lipschitz in θ . Thus, there exists a constant L_K independent of two numbers, θ_1 and θ_2 such that, $\|K(\theta_1) - K(\theta_2)\| \leq L_K \|\theta_1 - \theta_2\|$.
3. The function θ is Lipschitz in p . Thus, there exists a constant L_θ independent of two numbers, p_1 and p_2 such that, $\|\theta(p_1) - \theta(p_2)\| \leq L_\theta \|p_1 - p_2\|$.
4. The function $\theta(p)$ is monotone increasing in p , $\theta'(p) > 0$.
5. The specific storage $S_s = 0$.
6. The derivative $\partial_p K$ is bounded above, $|\partial_p K| \leq C$.

In this section, K is considered to be a function of p and not specifically of $\theta(p)$. The estimates given are simpler than in the previous section. As a consequence, we consider a discrete time scheme.

Again consider the variational formulation of Richards' equation given in equations (3.13)-(3.15). The discrete time numerical scheme considered here is that given in equations (3.33)-(3.35).

Before analyzing this case, we give without proof a lemma proven in [4] which will be used in the following arguments,

Lemma 4.3 Suppose that u^n , u^{n-1} , v^n and v^{n-1} are real numbers. Suppose also that $\theta : \mathbb{R} \mapsto \mathbb{R}$ is such that $0 \leq \theta' \leq Q < \infty$ and $|\theta''| \leq Q$ for some constant Q . Then,

$$d_t(\theta(u) - \theta(v))^n(u^n - v^n) = d_t \left(\int_v^u [\theta(\mu) - \theta(v)] d\mu \right)^n - E, \quad (4.60)$$

where

$$E \leq C' \{(u^n - v^n)^2 + (u^{n-1} - v^{n-1})^2 + (\Delta t^n)^2\}, \quad (4.61)$$

for some C' depending on Q and $|d_t u|$.

The main result of this section is as follows.

Theorem 4.2 Let $(P^n, \tilde{\mathbf{U}}^n, \mathbf{U}^n) \in (W_h, \mathbf{V}_h, \mathbf{V}_h^N)$ satisfy the equations (3.33)-(3.35) for each time step $n, n = 1, \dots, N$. Then,

$$\|P^N - p^N\| + \left(\sum_{n=1}^N K_* \|\tilde{\mathbf{U}}^n - \tilde{\mathbf{u}}^n\|^2 \Delta t^n \right)^{1/2} \leq C(h^{k+1} + \Delta t).$$

Proof Subtracting equations (3.13)-(3.15) from equations (3.33)-(3.35) gives the error equations,

$$\begin{aligned} (d_t(\theta(P) - \theta(p))^n, w) + (\nabla \cdot (\mathbf{U}^n - \Pi \mathbf{u}^n), w) \\ = -(d_t\theta(p^n) - \frac{\partial \theta(p^n)}{\partial t}, w), \end{aligned} \quad (4.62)$$

$$(\tilde{\mathbf{U}}^n - \hat{\tilde{\mathbf{u}}}^n, \mathbf{v}) = (P^n - \hat{p}^n, \nabla \cdot \mathbf{v}), \quad (4.63)$$

$$\begin{aligned} (\mathbf{U}^n - \Pi \mathbf{u}^n, \mathbf{v}) &= (\mathbf{u}^n - \Pi \mathbf{u}^n, \mathbf{v}) + (K(p^n)(\hat{\tilde{\mathbf{u}}}^n - \tilde{\mathbf{u}}^n), \mathbf{v}) \\ &\quad + (K(P^n)(\tilde{\mathbf{U}}^n - \hat{\tilde{\mathbf{u}}}^n), \mathbf{v}) + ((K(P^n) - K(p^n))\hat{\tilde{\mathbf{u}}}^n, \mathbf{v}). \end{aligned} \quad (4.64)$$

Let, $\gamma^n = P^n - \hat{p}^n$, $\boldsymbol{\eta}^n = \tilde{\mathbf{U}}^n - \hat{\tilde{\mathbf{u}}}^n$ and $\boldsymbol{\xi}^n = \mathbf{U}^n - \Pi \mathbf{u}^n$. Let $w = \gamma^n$ in (4.62), $\mathbf{v} = \boldsymbol{\xi}^n$ in (4.63) and $\mathbf{v} = \boldsymbol{\eta}^n$ in (4.64), giving,

$$\begin{aligned} (d_t(\theta(P^n) - \theta(p^n)), P^n - p^n) + (\nabla \cdot \boldsymbol{\xi}^n, \gamma^n) \\ = (d_t(\theta(P) - \theta(p))^n, \hat{p}^n - p^n) - (d_t\theta(p^n) - \frac{\partial \theta(p^n)}{\partial t}, \gamma^n), \\ (\boldsymbol{\eta}^n, \boldsymbol{\xi}^n) = (\gamma^n, \nabla \cdot \boldsymbol{\xi}^n), \end{aligned} \quad (4.65)$$

$$\begin{aligned} (\boldsymbol{\xi}^n, \boldsymbol{\eta}^n) &= +(\mathbf{u}^n - \Pi \mathbf{u}^n, \boldsymbol{\eta}^n) + (K(p^n)(\hat{\tilde{\mathbf{u}}}^n - \tilde{\mathbf{u}}^n), \boldsymbol{\eta}^n) + (K(P^n)\boldsymbol{\eta}^n, \boldsymbol{\eta}^n) \\ &\quad + ((K(P) - K(p))\hat{\tilde{\mathbf{u}}}^n, \boldsymbol{\eta}^n). \end{aligned} \quad (4.66)$$

Combine these equations to get,

$$\begin{aligned} (d_t(\theta(P) - \theta(p))^n, P^n - p^n) + (K(P^n)\boldsymbol{\eta}^n, \boldsymbol{\eta}^n) \\ = (d_t(\theta(P) - \theta(p))^n, \hat{p}^n - p^n) - (d_t\theta(p^n) - \frac{\partial \theta(p^n)}{\partial t}, \gamma^n) \\ - (\mathbf{u}^n - \Pi \mathbf{u}^n, \boldsymbol{\eta}^n) - (K(p^n)(\hat{\tilde{\mathbf{u}}}^n - \tilde{\mathbf{u}}^n), \boldsymbol{\eta}^n) \\ - ((K(P^n) - K(p^n))\hat{\tilde{\mathbf{u}}}^n, \boldsymbol{\eta}^n). \end{aligned} \quad (4.67)$$

By the Mean Value Theorem,

$$((K(P^n) - K(p^n))\hat{\mathbf{u}}^n, \boldsymbol{\eta}^n) = (K'(z)(P^n - p^n)\hat{\mathbf{u}}^n, \boldsymbol{\eta}^n),$$

where $z \in (P^n, p^n)$.

Consider the first term on the left-hand side of equation (4.67) and apply Lemma 4.3 to give,

$$\begin{aligned} (d_t(\theta(P) - \theta(p))^n, P^n - p^n) &\geq \int_{\Omega} \left(d_t \int_{p^n}^{P^n} (\theta(\mu) - \theta(p^n)) d\mu \right) dx \\ &\quad - C\{\|P^n - p^n\|^2 + \|P^{n-1} - p^{n-1}\|^2 \\ &\quad + (\Delta t^n)^2\}. \end{aligned} \quad (4.68)$$

Using the integral form of the remainder in Taylor's Theorem and the Schwarz inequality, the time discretization error term is bounded by,

$$\begin{aligned} |(d_t\theta(p^n) - \frac{\partial}{\partial t}\theta(p^n), \gamma^n)| &\leq C\|\gamma^n\|^2 + C\|d_t\theta(p^n) - \frac{\partial}{\partial t}\theta(p^n)\|^2 \\ &\leq C\|\gamma^n\|^2 + C\|\frac{1}{\Delta t^n} \int_{t^{n-1}}^{t^n} \frac{\partial^2 \theta(p)}{\partial t^2} (t - t^{n-1}) dt\|^2 \\ &\leq C\|\gamma^n\|^2 + C\Delta t^n \|\frac{\partial^2 \theta(p)}{\partial t^2}\|_{L^2((t^{n-1}, t^n); L^2)}^2. \end{aligned} \quad (4.69)$$

Combining the above bounds with equation (4.67),

$$\begin{aligned} &\int_{\Omega} \left(d_t \int_{p^n}^{P^n} (\theta(\mu) - \theta(p^n)) d\mu \right) dx + (K(P^n)\boldsymbol{\eta}^n, \boldsymbol{\eta}^n) \\ &\leq (d_t(\theta(P) - \theta(p))^n, \hat{p}^n - p^n) + \epsilon\|\boldsymbol{\eta}^n\|^2 \\ &\quad + C\{\|\Pi \mathbf{u}^n - \mathbf{u}^n\|^2 + \|\tilde{\mathbf{u}}^n - \hat{\mathbf{u}}^n\|^2 + \|P^n - p^n\|^2 + \|P^{n-1} - p^{n-1}\|^2 \\ &\quad + C\Delta t^n \|\frac{\partial^2 \theta(p)}{\partial t^2}\|_{L^2((t^{n-1}, t^n); L^2)}^2 + (\Delta t^n)^2\}. \end{aligned} \quad (4.70)$$

Taking ϵ small enough, the $\epsilon\|\boldsymbol{\eta}^n\|^2$ term can be brought to the left-hand side.

Multiply by Δt^n and sum on $n = 1, \dots, N$. The first term on the left-hand side collapses giving,

$$\begin{aligned} \sum_{n=1}^N \Delta t^n \int_{\Omega} \left(d_t \int_{p^n}^{P^n} (\theta(\mu) - \theta(p^n)) d\mu \right) dx &= \int_{\Omega} \int_{p^N}^{P^N} (\theta(\mu) - \theta(P^N)) d\mu dx \\ &\quad - \int_{\Omega} \int_{p^0}^{P^0} (\theta(\mu) - \theta(p^0)) d\mu dx. \end{aligned} \quad (4.71)$$

Applying summation by parts, the first term on the right-hand side becomes,

$$\begin{aligned}
& \sum_{n=1}^N (d_t(\theta(P^n) - \theta(p^n)), \hat{p}^n - p^n) \Delta t^n \\
&= (\theta(P^N) - \theta(p^N), \hat{p}^N - p^N) - (\theta(P^0) - \theta(p^0), \hat{p}^1 - p^1) \\
&\quad - \sum_{n=1}^N (\theta(P^n) - \theta(p^n), d_t(\hat{p}^{n+1} - p^{n+1})) \frac{\Delta t^{n+1}}{\Delta t^n} \Delta t^n \\
&\leq \epsilon \|P^N - p^N\|^2 + C \|\hat{p}^N - p^N\|^2 + C \|P^0 - p^0\|^2 + C \|\hat{p}^1 - p^1\|^2 \\
&\quad + \sum_{n=1}^N C \{ \|P^n - p^n\|^2 + \|\frac{\partial(\hat{p}^n - p^n)}{\partial t}\|^2 \} \Delta t^n, \tag{4.72}
\end{aligned}$$

where the assumption (3.32) has been used.

Since θ' is bounded above and below by positive constants, there exists a constant Q such that,

$$Q^{-1}(v - u)^2 \leq \int_u^v (\theta(\mu) - \theta(v)) d\mu \leq Q(v - u)^2. \tag{4.73}$$

Combining these bounds gives the following estimate,

$$\begin{aligned}
& \|P^N - p^N\|^2 + \sum_{n=1}^N K_* \|\boldsymbol{\eta}^n\|^2 \Delta t^n \\
&\leq C \|P^0 - p^0\|^2 + C \|\hat{p}^N - p^N\|^2 + \|\hat{p}^1 - p^1\|^2 \\
&\quad + C \sum_{n=1}^N \{ \|P^n - p^n\|^2 + \|\frac{\partial(\hat{p}^n - p^n)}{\partial t}\|^2 + \|\Pi \mathbf{u}^n - \mathbf{u}^n\|^2 + \|\tilde{\mathbf{u}}^n - \hat{\mathbf{u}}^n\|^2 \} \Delta t^n \\
&\quad + (\Delta t^n)^2. \tag{4.74}
\end{aligned}$$

Take $P^0 = \hat{p}^0$ and apply Gronwall's Lemma 3.1 to equation (4.74) to remove the first term in the sum on the right-hand side. Taking approximation properties of the L^2 and Π projections results in,

$$\|P^N - p^N\|^2 + \sum_{n=1}^N K_* \|\hat{\mathbf{u}} - \tilde{\mathbf{U}}^n\|^2 \Delta t^n \leq C(h^{2(k+1)} + (\Delta t)^2), \tag{4.75}$$

where k is the order of the approximating space. \square

Thus, in the case of strictly partially saturated flow, convergence for both hydraulic head and velocity is optimal.

4.3 Unsaturated to Fully Saturated Flow

In this section, the case of flow through unsaturated to fully saturated soil is considered. In this situation, K can be zero, and Richards' equation is degenerate. The general technique of Arbogast, Wheeler and Zhang [6] is followed for this analysis.

We write Richards' equation in the following way,

$$\frac{\partial \theta(p)}{\partial t} - \nabla \cdot (K(x)k(\theta(p))\nabla p) = f, \quad (4.76)$$

where in the case of fully saturated flow, we neglect the specific storage term. We allow for the case that $k(\theta(p)) = 0$, a condition arising when the porous media is at residual saturation. Note here that we assume the relative permeability is only a function of hydraulic head. The results given below can be generalized to the case where relative permeability also depends on position.

The following analysis will employ the Kirchhoff transformation,

$$R(p) = \int_0^p k(\theta(\wp))d\wp, \quad (4.77)$$

with gradient,

$$\nabla R(p) = k(\theta(p))\nabla p. \quad (4.78)$$

Defining $\tilde{\mathbf{u}} = -\nabla R$ and $\mathbf{u} = K(x)\tilde{\mathbf{u}}$, equation (4.76) can be written as the following equivalent system of equations,

$$\frac{\partial \theta(p)}{\partial t} + \nabla \cdot \mathbf{u} = f, \quad (4.79)$$

$$\tilde{\mathbf{u}} = -\nabla R, \quad (4.80)$$

$$\mathbf{u} = K(x)\tilde{\mathbf{u}}. \quad (4.81)$$

Alt and Luckhaus [3] state the following regularity results for the above equation,

$$\theta(p) \in L^\infty(J; L^1(\Omega)), \quad (4.82)$$

$$\partial_t \theta(p) \in L^2(J; H^{-1}(\Omega)), \quad (4.83)$$

$$\tilde{\mathbf{u}} \in L^2(J; (L^2(\Omega))^d). \quad (4.84)$$

Since $\partial_t \theta$ is only in $L^2(J; H^{-1}(\Omega))$, a variational formulation of the problem would require trial functions for equation (4.79) to be taken in $H^1(\Omega)$. In order to relax this requirement, an alternate time integrated variational formulation developed by Arbogast, Wheeler and Zhang [6] is considered.

For this formulation, we need to integrate θ in time, but equation (4.82) does not guarantee that $\theta(p)$ exists pointwise everywhere in time. However, we know that physically θ is defined at every time and we assume that $\theta(p) \in L^\infty(J; L^\infty(\Omega))$ so that $\theta(p)$ exists pointwise for each t . Therefore, (4.79) can be integrated to get,

$$\theta(p(\cdot, t)) + \nabla \cdot \int_0^t \mathbf{u} d\tau = \int_0^t f d\tau + \theta(p^0). \quad (4.85)$$

Since physically, the normal components of the flow flux are continuous, $\nabla \cdot \mathbf{u} \in L^2(\Omega)$. Thus, the integral $\int_0^t \mathbf{u} d\tau$ is in $L^2(J; H(\Omega; \text{div}))$, and the following variational formulation can be defined,

$$(\theta(p), w) + (\nabla \cdot \int_0^t \mathbf{u} d\tau, w) = (\int_0^t f d\tau, w) + (\theta(p^0), w), \quad w \in W, \quad (4.86)$$

$$(\tilde{\mathbf{u}}, \mathbf{v}) = (R(p), \nabla \cdot \mathbf{v}) - (R(p_D), \mathbf{v} \cdot \mathbf{n})_{\Gamma^D}, \quad \mathbf{v} \in \mathbf{V}^0, \quad (4.87)$$

$$(\mathbf{u}, \mathbf{v}) = (K(x)\tilde{\mathbf{u}}, \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}. \quad (4.88)$$

The continuous time numerical scheme is to find $(P, \tilde{\mathbf{U}}, \mathbf{U}) \in (W_h, \mathbf{V}_h, \mathbf{V}_h^N)$ satisfying,

$$(\theta(P), w) + (\nabla \cdot \int_0^t \mathbf{U} d\tau, w) = (\int_0^t f d\tau, w) + (\theta(P^0), w), \quad w \in W_h, \quad (4.89)$$

$$(\tilde{\mathbf{U}}, \mathbf{v}) = (R(P), \nabla \cdot \mathbf{v}) - (R(p_D), \mathbf{v} \cdot \mathbf{n})_{\Gamma^D}, \quad \mathbf{v} \in \mathbf{V}_h^0, \quad (4.90)$$

$$(\mathbf{U}, \mathbf{v}) = (K(x)\tilde{\mathbf{U}}, \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_h. \quad (4.91)$$

Theorem 4.3 For the numerical scheme given by equations (4.89)-(4.91), the following bounds hold,

$$\begin{aligned} & \int_0^T (\theta(P) - \theta(p), e^{-rt}(\widehat{R(P)} - \widehat{R(p)})) + \|K^{1/2} \int_0^T \boldsymbol{\eta} ds\|^2 \\ & \leq C \{ e^{-rT} \|\int_0^T K^{1/2}(\tilde{\mathbf{u}} - \hat{\mathbf{u}}) d\tau\|^2 + \int_0^T e^{-rt} \|K^{1/2}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})\|^2 dt \\ & \quad + e^{-rT} \|\int_0^T (\mathbf{u} - \Pi \mathbf{u}) d\tau\|^2 + \int_0^T e^{-rt} \|\mathbf{u} - \Pi \mathbf{u}\|^2 dt \}. \end{aligned} \quad (4.92)$$

Proof Making use of the L^2 and Π projections, and then subtracting equations (4.86)-(4.88) from (4.89)-(4.91) gives the following error equations,

$$(\theta(P) - \theta(p), w) + (\nabla \cdot \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau, w) = (\theta(P^0) - \theta(p^0), w), \quad w \in W_h, \quad (4.93)$$

$$(\tilde{\mathbf{U}} - \hat{\mathbf{u}}, \mathbf{v}) = (\widehat{R(P)} - \widehat{R(p)}, \nabla \cdot \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_h^0, \quad (4.94)$$

$$(\mathbf{U} - \Pi \mathbf{u}, \mathbf{v}) = (K(x)(\tilde{\mathbf{U}} - \tilde{\mathbf{u}}), \mathbf{v}) + (\mathbf{u} - \Pi \mathbf{u}, \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_h. \quad (4.95)$$

For some constant r defined later, let $w = e^{-rt}(\widehat{R(P)} - \widehat{R(p)}) \in W_h$ in (4.93) resulting in,

$$\begin{aligned} & (\theta(P) - \theta(p), e^{-rt}(\widehat{R(P)} - \widehat{R(p)})) + (\nabla \cdot \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau, e^{-rt}(\widehat{R(P)} - \widehat{R(p)})) \\ & = (\theta(P^0) - \theta(p^0), e^{-rt}(\widehat{R(P)} - \widehat{R(p)})). \end{aligned} \quad (4.96)$$

In (4.94), let $\mathbf{v} = e^{-rt} \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau$ to give,

$$(\tilde{\mathbf{U}} - \hat{\mathbf{u}}, e^{-rt} \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau) = (\widehat{R(P)} - \widehat{R(p)}, \nabla \cdot e^{-rt} \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau). \quad (4.97)$$

Lastly, integrate (4.95) from 0 to t holding \mathbf{v} fixed and multiply by e^{rt} . Then, let $\mathbf{v} = \tilde{\mathbf{U}} - \hat{\mathbf{u}}$ so that,

$$\begin{aligned} & (e^{-rt} \int_0^t (\mathbf{U} - \Pi \mathbf{u}) d\tau, \tilde{\mathbf{U}} - \hat{\mathbf{u}}) \\ & = (e^{-rt} \int_0^t K(x)(\tilde{\mathbf{U}} - \hat{\mathbf{u}}) d\tau, \tilde{\mathbf{U}} - \hat{\mathbf{u}}) + (e^{-rt} \int_0^t (\mathbf{u} - \Pi \mathbf{u}) d\tau, \tilde{\mathbf{U}} - \hat{\mathbf{u}}). \end{aligned} \quad (4.98)$$

Combining the above three equations results in,

$$\begin{aligned} & (\theta(P) - \theta(p), e^{-rt}(\widehat{R(P)} - \widehat{R(p)})) + \left(e^{-rt} \int_0^t K(x)(\tilde{\mathbf{U}} - \hat{\mathbf{u}}), \tilde{\mathbf{U}} - \hat{\mathbf{u}} \right) \\ & = \left(e^{-rt} \int_0^t K(x)(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \tilde{\mathbf{U}} - \hat{\mathbf{u}} \right) + \left(e^{-rt} \int_0^t (\Pi \mathbf{u} - \mathbf{u}), \tilde{\mathbf{U}} - \hat{\mathbf{u}} \right) \\ & \quad + (\theta(P^0) - \theta(p^0), e^{-rt}(\widehat{R(P)} - \widehat{R(p)})). \end{aligned} \quad (4.99)$$

Let $\boldsymbol{\eta} = \tilde{\mathbf{U}} - \hat{\mathbf{u}}$ and integrate (4.99) in time from 0 to T . Then, consider the second left-hand side term of (4.99),

$$\frac{d}{dt} \left(K(x) \int_0^t \boldsymbol{\eta} ds, \int_0^t \boldsymbol{\eta} ds \right) = 2 \left(K(x) \boldsymbol{\eta}, \int_0^t \boldsymbol{\eta} ds \right). \quad (4.100)$$

So, $(e^{-rt} \int_0^t K(x) \boldsymbol{\eta}, \boldsymbol{\eta}) = \frac{1}{2} e^{-rt} \frac{d}{dt} \|K^{1/2} \int_0^t \boldsymbol{\eta} ds\|^2$. Thus, by integration by parts,

$$\begin{aligned} \int_0^T \left(K(x) \boldsymbol{\eta}, e^{-rt} \int_0^t \boldsymbol{\eta} ds \right) & = \frac{1}{2} \int_0^T r e^{-rt} \left\| K^{1/2} \int_0^t \boldsymbol{\eta} ds \right\|^2 dt \\ & \quad + e^{-rT} \frac{1}{2} \left\| K^{1/2} \int_0^T \boldsymbol{\eta} ds \right\|^2. \end{aligned} \quad (4.101)$$

Therefore, (4.99) integrated in time is,

$$\int_0^T \left(\theta(P) - \theta(p), e^{-rt}(\widehat{R(P)} - \widehat{R(p)}) \right) + \frac{1}{2} \int_0^T r e^{-rt} \left\| K^{1/2} \int_0^t \boldsymbol{\eta} ds \right\|^2 dt$$

$$\begin{aligned}
& + e^{-rT} \frac{1}{2} \left\| K^{1/2} \int_0^T \boldsymbol{\eta} ds \right\|^2 \\
& = \int_0^T \left(e^{-rt} \int_0^t K(x)(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \boldsymbol{\eta} \right) + \int_0^T \left(e^{-rt} \int_0^t e^{r\tau} (\Pi \mathbf{u} - \mathbf{u}), \boldsymbol{\eta} \right) \\
& \quad + \int_0^T (\theta(P^0) - \theta(p^0), e^{-rt}(\widehat{R(P)} - \widehat{R(p)})). \tag{4.102}
\end{aligned}$$

Now, by integration by parts,

$$\begin{aligned}
& \int_0^T \left(e^{-rt} \int_0^t K(x)(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \boldsymbol{\eta} \right) dt \\
& = r \int_0^T \left(e^{-rt} \int_0^t K(x)(\tilde{\mathbf{u}} - \hat{\mathbf{u}}) d\tau, \int_0^t \boldsymbol{\eta} d\tau \right) dt - \int_0^T \left(e^{-rt} K(x)(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \int_0^t \boldsymbol{\eta} d\tau \right) dt \\
& \quad + \left(e^{-rT} \int_0^T K(x)(\tilde{\mathbf{u}} - \hat{\mathbf{u}}), \int_0^T \boldsymbol{\eta} d\tau \right) \\
& \leq Cr \int_0^T e^{-rt} \left\| K^{1/2} \int_0^t (\tilde{\mathbf{u}} - \hat{\mathbf{u}}) d\tau \right\|^2 dt + \epsilon r \int_0^T e^{-rt} \left\| K^{1/2} \int_0^t \boldsymbol{\eta} d\tau \right\|^2 dt \\
& \quad + C \int_0^T e^{-rt} \|K^{1/2}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})\|^2 dt + \epsilon \int_0^T e^{-rt} \left\| K^{1/2} \int_0^t \boldsymbol{\eta} d\tau \right\|^2 dt \\
& \quad + C e^{-rT} \left\| \int_0^T K^{1/2}(\tilde{\mathbf{u}} - \hat{\mathbf{u}}) d\tau \right\|^2 + \epsilon e^{-rT} \left\| K^{1/2} \int_0^T \boldsymbol{\eta} d\tau \right\|^2. \tag{4.103}
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \int_0^T \left(e^{-rt} \int_0^t (\Pi \mathbf{u} - \mathbf{u}), \boldsymbol{\eta} \right) \\
& \leq Cr \int_0^T e^{-rt} \left\| \int_0^t (\mathbf{u} - \Pi \mathbf{u}) d\tau \right\|^2 dt + \tilde{\epsilon} r \int_0^T e^{-rt} \left\| \int_0^t \boldsymbol{\eta} d\tau \right\|^2 dt \\
& \quad + C \int_0^T e^{-rt} \|\mathbf{u} - \Pi \mathbf{u}\|^2 dt + \tilde{\epsilon} \int_0^T e^{-rt} \left\| \int_0^t \boldsymbol{\eta} d\tau \right\|^2 dt \\
& \quad + C e^{-rT} \left\| \int_0^T (\mathbf{u} - \Pi \mathbf{u}) d\tau \right\|^2 + \tilde{\epsilon} e^{-rT} \left\| \int_0^T \boldsymbol{\eta} d\tau \right\|^2. \tag{4.104}
\end{aligned}$$

Take the initial approximation, P^0 , such that, $(\theta(P^0) - \theta(p^0), w) = 0, \forall w \in W_h$. This is done by computing $P^0 = \theta^{-1}(\widehat{\theta(p^0)})$, where $\theta(\hat{p}^0)|_{E_i} = \frac{1}{m(E_i)} \int_{E_i} \theta(p^0)$.

Thus, combining equations (4.103)-(4.104) with equation (4.102), taking ϵ and $\tilde{\epsilon}$ small enough and choosing r large enough to exactly cancel the second left-hand side term in (4.102) gives,

$$\int_0^T (\theta(P) - \theta(p), e^{-rt}(\widehat{R(P)} - \widehat{R(p)})) + \left\| K^{1/2} \int_0^T \boldsymbol{\eta} ds \right\|^2$$

$$\begin{aligned}
&\leq C \{ e^{-rT} \left\| \int_0^T K^{1/2}(\tilde{\mathbf{u}} - \hat{\mathbf{u}}) d\tau \right\|^2 + \int_0^T e^{-rt} \|K^{1/2}(\tilde{\mathbf{u}} - \hat{\mathbf{u}})\|^2 dt \\
&\quad + e^{-rT} \left\| \int_0^T (\mathbf{u} - \Pi \mathbf{u}) d\tau \right\|^2 + \int_0^T e^{-rt} \|\mathbf{u} - \Pi \mathbf{u}\|^2 dt \}. \quad (4.105)
\end{aligned}$$

□

The theorem just proven bounds the error in the numerical flux by approximation bounds. Thus, once the regularity of the solution is known, the error in the flux will have the same asymptotic behavior as approximation in the discrete space.

The next estimate gives a bound in the H_{-1} norm of the numerical approximation to hydraulic head in the case that $\partial\theta/\partial p > 0$. The above result is used to derive this result.

Let $\psi \in H_0^1(\Omega)$. Then, equation (4.93) and the definition of the L^2 projection imply,

$$\begin{aligned}
(\theta(P) - \theta(p), \psi) &= (\theta(P) - \theta(p), \psi - \hat{\psi}) + (\theta(P) - \theta(p), \hat{\psi}) \\
&= (\theta(P) - \theta(p), \psi - \hat{\psi}) - \left(\nabla \cdot \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau, \hat{\psi} \right) \\
&= (\theta(P) - \theta(p), \psi - \hat{\psi}) + \left(\Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau, \nabla \psi \right), \quad (4.106)
\end{aligned}$$

where we have used integration by parts, the definition of the Π projection and have again choosen P^0 so that the initial term is 0.

Now, assuming θ is Lipschitz continuous, and again using the definition of the L^2 projection,

$$\begin{aligned}
(\theta(P) - \theta(p), \psi - \hat{\psi}) &= (\theta(P) - \theta(\hat{p}), \psi - \hat{\psi}) + (\theta(\hat{p}) - \theta(p), \psi - \hat{\psi}) \\
&\leq 0 + Ch \|\hat{p} - p\| \|\psi\|_{H_1}. \quad (4.107)
\end{aligned}$$

Also,

$$\left(\Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau, \nabla \psi \right) \leq \left\| \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau \right\| \|\psi\|_{H_1}. \quad (4.108)$$

Integrating equation (4.95) from 0 to t holding \mathbf{v} fixed and then taking $\mathbf{v} = \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau$, results in,

$$\left\| \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau \right\|^2 = \left(K(x) \int_0^t (\tilde{\mathbf{U}} - \tilde{\mathbf{u}}) d\tau, \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau \right)$$

$$\begin{aligned}
& + \left(\int_0^t (\mathbf{u} - \Pi \mathbf{u}) d\tau, \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau \right) \\
& \leq C \left\| K^{1/2} \int_0^t (\tilde{\mathbf{U}} - \tilde{\mathbf{u}}) d\tau \right\|^2 + C \left\| \int_0^t (\mathbf{u} - \Pi \mathbf{u}) d\tau \right\|^2 \\
& \quad + \epsilon \left\| \Pi \int_0^t (\mathbf{U} - \mathbf{u}) d\tau \right\|^2. \tag{4.109}
\end{aligned}$$

Combining equations (4.106)-(4.109) and recalling the definition of the H_{-1} norm, equation (3.1), gives,

$$\begin{aligned}
& \|\theta(P) - \theta(p)\|_{H_{-1}} \\
& \leq C \{h\|\hat{p} - p\| + \left\| K^{1/2} \int_0^t (\tilde{\mathbf{U}} - \tilde{\mathbf{u}}) d\tau \right\| + \left\| \int_0^t (\mathbf{u} - \Pi \mathbf{u}) d\tau \right\|\}. \tag{4.110}
\end{aligned}$$

For a given time t , apply the mean value theorem to write,

$$\theta(P(\cdot, t)) - \theta(p(\cdot, t)) = \theta'(w(t))(P(\cdot, t) - p(\cdot, t)) \geq C(P - p). \tag{4.111}$$

Thus,

$$\begin{aligned}
& \|P - p\|_{H_{-1}} \\
& \leq C \{h\|\hat{p} - p\| + \left\| K^{1/2} \int_0^t (\tilde{\mathbf{U}} - \tilde{\mathbf{u}}) d\tau \right\| + \left\| \int_0^t (\mathbf{u} - \Pi \mathbf{u}) d\tau \right\|\}. \tag{4.112}
\end{aligned}$$

So, with Theorem 4.3, the error in hydraulic head is bounded in terms of approximation error.

Chapter 5

Two-Level Methods for Nonlinear Parabolic Equations

The analysis in the previous chapter applies to discretizing the full nonlinear problem on a computational grid of cell diameter h . However, due to the highly nonlinear nature of θ and K , solving the resulting discrete nonlinear system is computationally very expensive. Thus, alternative schemes which get around solving the full nonlinear mixed formulation of Richards' equation are attractive.

One alternative approach is to consider linearizing the equation before discretization. For this approach, one would solve the full nonlinear problem on a coarse grid with a small amount of unknowns, then use that coarse grid solution to linearize the problem on a fine grid. This idea of using a two level scheme for nonlinear problems was first developed by Xu [63, 64] who applied it to nonlinear elliptic equations with Galerkin finite elements and extended by Dawson and Wheeler [22] to the expanded mixed method applied to nonlinear parabolic equations.

Xu showed optimal H_1 convergence for both the coarse and fine grids. Dawson and Wheeler showed optimal H_1 and L^2 estimates for the coarse and fine grids, and for the case of the lowest order Raviart-Thomas-Nedelec space, they showed superconvergence for the coarse grid in both norms.

In this work, we will show superconvergence results in certain discrete norms on both grids for a finite difference scheme applied to the nonlinear heat equation. This is a first step in trying to apply the two-level technique to Richards' equation. After completing this analysis. We show convergence results for a two-level scheme with the expanded mixed method applied to Richards' equation. For this scheme, the equation is not fully linearized on the fine grid. To fully linearize the equation would require giving up a mass conserving scheme. Previous work has shown that solutions become inaccurate when mass balance is lost. We thus leave the time term nonlinear on the fine grid and just consider linearizing the hydraulic conductivity term.

Two quasi-uniform triangulations of Ω are considered, a coarse triangulation with mesh size H denoted by \mathcal{T}_H , and a fine triangulation with mesh size h denoted by \mathcal{T}_h . We assume that \mathcal{T}_h is a refinement of \mathcal{T}_H . Both these triangulations consist of rectangles in two dimensions and parallelepipeds in three dimensions.

5.1 A Two-Level Finite Difference Scheme

We begin with a finite difference scheme for the nonlinear heat equation,

$$\frac{\partial p}{\partial t} - \nabla \cdot K(p) \nabla p = f, \quad (5.1)$$

$$-K(p) \nabla p \cdot \mathbf{n} = 0. \quad (5.2)$$

For simplicity we consider homogeneous Neumann boundary conditions. It is straightforward to extend the following results to nonhomogeneous conditions. The following assumptions are made.

1. The tensor K is symmetric and positive definite.
2. The tensor K is bounded, i.e. there exist positive constants, K_* and K^* such that for $z \in \mathbb{R}^d$,

$$K_* \|z\|^2 \leq z^t K z \leq K^* \|z\|.$$

3. Each element of K is twice continuously differentiable in space and time with derivatives up to second order bounded above by K^* .
4. The tensor $K(p)$ is Lipschitz continuous in p .

5.1.1 A Coarse Grid Nonlinear Finite Difference Scheme

In this section we develop and give convergence estimates for a nonlinear cell-centered finite difference scheme on the coarse grid. For simplicity we consider two dimensions and note that extensions to three dimensions are straightforward.

Definition of the Scheme

A variational formulation for (5.1)-(5.2) at time t^n is to find $(p^n, \tilde{\mathbf{u}}^n, \mathbf{u}^n) \in (W \times V \times V^0)$ satisfying

$$(\partial_t p^n, w) + (\nabla \cdot \mathbf{u}^n, w) = (f^n, w), \quad \forall w \in W, \quad (5.3)$$

$$(\tilde{\mathbf{u}}^n, \mathbf{v}) = (p^n, \nabla \cdot \mathbf{v}), \quad \forall \mathbf{v} \in V^0, \quad (5.4)$$

$$(\mathbf{u}^n, \mathbf{v}) = (K(p^n)\tilde{\mathbf{u}}^n, \mathbf{v}), \quad \forall \mathbf{v} \in V. \quad (5.5)$$

Cell-centered finite difference approximations $P_H^n \in W_H$, $\tilde{\mathbf{U}}_H^n \in V_H$ and $\mathbf{U}_H^n \in V_H^0$ to the functions $p(t^n, \cdot)$, $\tilde{\mathbf{u}}(t^n, \cdot)$ and $\mathbf{u}(t^n, \cdot)$, respectively, are chosen for each $n = 1, \dots, N$, satisfying

$$(d_t P_H^n, w) + (\nabla \cdot \mathbf{U}_H^n, w) = (f^n, w), \quad \forall w \in W_H, \quad (5.6)$$

$$(\tilde{\mathbf{U}}_H^n, \mathbf{v})_{\text{TM}} = (P_H^n, \nabla \cdot \mathbf{v}), \quad \forall \mathbf{v} \in V_H^0, \quad (5.7)$$

$$(\mathbf{U}_H^n, \mathbf{v})_{\text{TM}} = (K(\mathcal{P}_H(P_H^n))\tilde{\mathbf{U}}_H^n, \mathbf{v})_{\text{T}}, \quad \forall \mathbf{v} \in V_H, \quad (5.8)$$

with $P_H^0 = \hat{p}_H(t^0, \cdot)$.

Recalling that the grid points are denoted by,

$$(x_{i+1/2}, y_{j+1/2}), \quad i = 0, \dots, N_x, \quad j = 0, \dots, N_y,$$

and midpoints by,

$$x_i = \frac{1}{2}(x_{i+1/2} + x_{i-1/2}), \quad i = 1, \dots, N_x,$$

$$y_j = \frac{1}{2}(y_{j+1/2} + y_{j-1/2}), \quad j = 1, \dots, N_y,$$

define $\mathcal{P}_H(p)$ from the values of p_{ij} for $i = 1, \dots, \hat{N}_x$ and $j = 1, \dots, \hat{N}_y$ as follows. For points (x, y) such that $x_i \leq x \leq x_{i+1}$, $i \in \{1, \dots, \hat{N}_x\}$ and $y_j \leq y \leq y_{j+1}$, $j \in \{1, \dots, \hat{N}_y\}$, take $\mathcal{P}_H(p)(x, y)$ to be the bilinear interpolant,

$$\begin{aligned} \mathcal{P}_H(p)(x, y) &= (p_{ij}(\frac{x_{i+1} - x}{x_{i+1} - x_i}) + p_{i+1j}(\frac{x - x_i}{x_{i+1} - x_i}))(\frac{y_{j+1} - y}{y_{j+1} - y_j}) \\ &\quad + (p_{ij+1}(\frac{x_{i+1} - x}{x_{i+1} - x_i}) + p_{i+1j+1}(\frac{x - x_i}{x_{i+1} - x_i}))(\frac{y - y_j}{y_{j+1} - y_j}). \end{aligned}$$

For $i = 1, \dots, \hat{N}_x - 1$, set

$$\mathcal{P}_H(p)(x_i, y_{1/2}) = \frac{(2H_1^y + H_2^y)p_{i1} - H_1^y p_{i2}}{H_1^y + H_2^y}.$$

This is a two point extrapolation, and by Taylor's theorem $|(\mathcal{P}_H(p) - p)(x_i, y_{1/2})| \leq CH^2$. For points (x, y) such that $x_i \leq x \leq x_{i+1}$ and $y_{1/2} \leq y \leq y_1$, define $\mathcal{P}_H(p)$ as the bilinear interpolant between $p_{i,1}$, $p_{i+1,1}$, $\mathcal{P}_H(p)(x_i, y_{1/2})$ and $\mathcal{P}_H(p)(x_{i+1}, y_{1/2})$. By interpolation theory $|\mathcal{P}_H(p) - p| \leq CH^2$ for these points. In a similar way define

$\mathcal{P}_H(p)$ for (x, y) such that $x_i \leq x \leq x_{i+1}$ and $y_{\hat{N}_y} \leq y \leq y_{\hat{N}_y+1/2}$ as well as for points (x, y) where $x_{1/2} \leq x \leq x_1$ or $x_{\hat{N}_x} \leq x \leq x_{\hat{N}_x+1/2}$ and $y_j \leq y \leq y_{j+1}$ for j such that $1 \leq j \leq \hat{N}_y$. Lastly, define $\mathcal{P}_H(p)$ at the corners of the domain. Here, three point extrapolation is used,

$$\begin{aligned}\mathcal{P}_H(p)(x_{1/2}, y_{1/2}) &= \mathcal{P}_H(p)_{1,1/2} + \mathcal{P}_H(p)_{1/2,1} - p_{1,1} \\ &= p_{1,1/2} + p_{1/2,1} - p_{1,1} + O(H^2).\end{aligned}$$

By Taylor's theorem, $|(\mathcal{P}_H(p) - p)(x_{1/2}, y_{1/2})| \leq CH^2$. For points (x, y) such that $x_{1/2} \leq x \leq x_1$ and $y_{1/2} \leq y \leq y_1$, define $\mathcal{P}_H(p)(x, y)$ as the bilinear interpolant of $\mathcal{P}_H(p)(x_{1/2}, y_{1/2})$, $\mathcal{P}_H(p)(x_{1/2}, y_1)$, $\mathcal{P}_H(p)(x_1, y_{1/2})$ and $p_{1,1}$ which is an $O(H^2)$ approximation to $p(x, y)$ within this "corner region". Similarly, define $\mathcal{P}_H(p)$ as an $O(H^2)$ approximation to p in the other three "corner" regions.

We have just proven the following lemma,

Lemma 5.1 If p is twice differentiable, then for $\mathcal{P}_H(p)$ defined above,

$$\|\mathcal{P}_H(p) - p\|_{L^\infty} \leq CH^2.$$

If a uniform mesh is used and K is a diagonal tensor, equations (5.6)-(5.8) reduce to a standard nonlinear finite difference procedure. Denoting P_H^n by P^n , in the interior of Ω ,

$$\begin{aligned}f_{ij}^n H^2 + P_{ij}^{n-1} \frac{H^2}{\Delta t^n} \\ = \frac{1}{2} [(K_{11}(\mathcal{P}_H(P^n))_{i+1/2j+1/2} + K_{11}(\mathcal{P}_H(P^n))_{i+1/2j-1/2})(P_{ij}^n - P_{i+1j}^n) \\ + (K_{11}(\mathcal{P}_H(P^n))_{i-1/2j+1/2} + K_{11}(\mathcal{P}_H(P^n))_{i-1/2j-1/2})(P_{ij}^n - P_{i-1j}^n) \\ + (K_{22}(\mathcal{P}_H(P^n))_{i+1/2j+1/2} + K_{22}(\mathcal{P}_H(P^n))_{i-1/2j+1/2})(P_{ij}^n - P_{ij+1}^n) \\ + (K_{22}(\mathcal{P}_H(P^n))_{i+1/2j-1/2} + K_{22}(\mathcal{P}_H(P^n))_{i-1/2j-1/2})(P_{ij}^n - P_{ij-1}^n)] \\ + \frac{H^2}{\Delta t^n} P_{ij}^n.\end{aligned}\tag{5.9}$$

Existence and uniqueness of a solution to this discrete nonlinear problem is given in the following theorem.

Theorem 5.1 For time t^n and Δt sufficiently small, there exists a unique solution to equations (5.6)-(5.8).

Proof We are seeking a unique solution to the nonlinear equation $F(P^n) = 0$, where $F(P^n) = b^n + P^n + \frac{\Delta t^n}{H^2} A(P^n) P^n$. Here, b^n is a vector whose entry corresponding to grid cell (x_i, y_j) is $-\frac{\Delta t^n}{H^2} \int_{\Omega_{ij}} f_{ij}^n - P_{ij}^{n-1}$, P is a vector whose ij th entry corresponds to the value of the scalar variable P_{ij}^n and A is a matrix function of P^n given by the stencil above. By Theorem 5.4.5 of Ortega and Reinbolt [52], if F is continuously differentiable and uniformly monotone on \mathbb{R}^n , then a unique solution to $F(P^n) = 0$ exists. It is easily verified that the F defined above is continuously differentiable. In order to prove that F is uniformly monotonic we note that uniform monotonicity is equivalent to positive definiteness of the Jacobian, $J = F'$, and that a real matrix J is positive definite if and only if its symmetric part, $(J + J^T)/2$, is positive definite [10, Lemma 3.1]. Furthermore, we know that if a matrix is strictly diagonal dominant with positive diagonal entries, then the eigenvalues of the matrix have positive real parts [10, Theorem 4.9]. Now, $J = I + \frac{\Delta t^n}{H^2} A(P^n) + \frac{\Delta t^n}{H^2} A'(P^n) P^n$. Thus, with $\frac{\Delta t^n}{H^2}$ sufficiently small, we have that the symmetric part of J has positive real eigenvalues and, hence, is positive definite, making J positive definite and F uniformly monotonic. \square

Preliminary Estimates

Before we show convergence estimates for this finite difference scheme, we show convergence for a related linear scheme. The arguments given below closely follow those of Arbogast, Wheeler and Yotov [8] except that we extend their work to time differenced time dependent problems.

Theorem 5.2 For each $n = 1, \dots, N$, let $(\underline{P}_H^n, \tilde{\underline{U}}_H^n, \underline{U}_H^n) \in (W_H \times V_H \times V_H^0)$ satisfy

$$(\nabla \cdot \underline{U}_H^n, w) = (b^n, w), \quad \forall w \in W_H, \quad (5.10)$$

$$(\tilde{\underline{U}}_H^n, \mathbf{v})_{\text{TM}} = (\underline{P}_H^n, \nabla \cdot \mathbf{v}), \quad \forall \mathbf{v} \in V_H^0, \quad (5.11)$$

$$(\underline{U}_H^n, \mathbf{v})_{\text{TM}} = (K(\mathcal{P}_H(p^n)) \tilde{\underline{U}}_H^n, \mathbf{v})_{\text{T}}, \quad \forall \mathbf{v} \in V_H, \quad (5.12)$$

with $b^n = f^n - \partial_t p^n$ and $\underline{P}_H^n = \hat{p}_H^n$. We further require the compatibility condition, $\int_{\Omega} \underline{P}_H^n = \int_{\Omega} \hat{p}^n$. Then,

$$\|\underline{U}_H^n - \mathbf{u}^n\|_{\text{TM}} + \|\tilde{\underline{U}}_H^n - \tilde{\mathbf{u}}^n\|_{\text{TM}} \leq CH^2, \quad (5.13)$$

$$\|\underline{P}_H^n - p^n\|_{\text{M}} \leq CH^2, \quad (5.14)$$

$$\|d_t \underline{P}_H^n - d_t p^n\|_{\text{M}} \leq C(H^2 + \Delta t). \quad (5.15)$$

We will make use of the following lemma proven in Arbogast, Wheeler and Yotov [8].

Lemma 5.2 For the lowest order RTN space on rectangles and for any $\mathbf{q} = (q^x, q^y) \in H^1(\Omega)$ and $E \in \mathcal{T}_k$,

$$\left\| \frac{\partial}{\partial x} (\Pi \mathbf{q})^x \right\|_{0,E} \leq \left\| \frac{\partial q^x}{\partial x} \right\|_{0,E}, \quad (5.16)$$

$$\left\| \frac{\partial}{\partial y} (\Pi \mathbf{q})^y \right\|_{0,E} \leq \left\| \frac{\partial q^y}{\partial y} \right\|_{0,E}. \quad (5.17)$$

In order to prove the above theorem, we will first prove two preliminary lemmas.

Lemma 5.3 Assume for each $n = 1, \dots, N$, that $p^n, \partial_t p^n \in W_3^4(\Omega)$. Then, there exist $\tilde{\mathbf{U}}^{*,n} \in V_H$, $P^{*,n} \in W_H$, $\mathbf{Z}^{*,n} \in V_H^n$, $\tilde{\mathbf{Z}}^{*,n} \in V_H$ and $W^{*,n} \in W_H$ such that

$$(\tilde{\mathbf{U}}^{*,n}, \mathbf{v})_{\text{TM}} = (P^{*,n}, \nabla \cdot \mathbf{v}), \quad \mathbf{v} \in V_H^0, \quad (5.18)$$

$$(\tilde{\mathbf{Z}}^{*,n}, \mathbf{v})_{\text{TM}} = (W^{*,n}, \nabla \cdot \mathbf{v}), \quad \mathbf{v} \in V_H^0, \quad (5.19)$$

$$(\mathbf{Z}^{*,n}, \mathbf{v})_{\text{TM}} = (K(p^n) \tilde{\mathbf{Z}}^{*,n}, \mathbf{v})_{\text{T}} + (\partial_t(K(p^n)) \tilde{\mathbf{U}}^{*,n}, \mathbf{v})_{\text{T}}, \quad \mathbf{v} \in V_H, \quad (5.20)$$

and there exists a constant C independent of H such that for all i, j ,

$$|P_{i,j}^{*,n} - p_{i,j}^n| \leq CH^2, \quad (5.21)$$

$$|W_{i,j}^{*,n} - \partial_t p_{i,j}^n| \leq CH^2, \quad (5.22)$$

$$|\tilde{U}_{x,i+1/2j}^{*,n} - \tilde{u}_{x,i+1/2j}^n| + |\tilde{U}_{y,ij+1/2}^{*,n} - \tilde{u}_{y,ij+1/2}^n| \leq CH^2, \quad (5.23)$$

$$|\tilde{Z}_{x,i+1/2j}^{*,n} - \partial_t \tilde{u}_{x,i+1/2j}^n| + |\tilde{Z}_{y,ij+1/2}^{*,n} - \partial_t \tilde{u}_{y,ij+1/2}^n| \leq CH^2, \quad (5.24)$$

$$|Z_{x,i+1/2j}^{*,n} - \partial_t u_{x,i+1/2j}^n| + |Z_{y,ij+1/2}^{*,n} - \partial_t u_{y,ij+1/2}^n| \leq CH^2. \quad (5.25)$$

Proof Arbogast, Wheeler and Yotov [8] present a lemma which gives the desired $P^{*,n}$ and $\tilde{\mathbf{U}}^{*,n}$ above. In order to derive (5.22) and (5.24), we apply a lemma due to

Weiser and Wheeler [62] to the pair $(\partial_t \tilde{\mathbf{u}}^n, \partial_t p^n)$ which, by definition, satisfies the two equations,

$$\begin{aligned}\nabla \cdot \partial_t \tilde{\mathbf{u}}^n &= F^n, \text{ in } \Omega, \\ \partial_t \tilde{\mathbf{u}}^n &= -\nabla \partial_t p^n, \text{ in } \Omega,\end{aligned}$$

where $F^n = \partial_t f^n + \partial_t p^n$. This result gives a $W^{*,n}$ satisfying (5.22) and through (5.19), $\tilde{\mathbf{Z}}^{*,n}$ satisfies (5.24) in the interior of Ω . Define $\tilde{\mathbf{Z}}$ on Γ by,

$$\begin{aligned}\tilde{\mathbf{Z}}_{x,i+1/2j}^{*,n} &= \partial_t \tilde{\mathbf{u}}_{x,i+1/2j}^n, \\ \tilde{\mathbf{Z}}_{y,ij+1/2}^{*,n} &= \partial_t \tilde{\mathbf{u}}_{y,ij+1/2}^n.\end{aligned}$$

Then, (5.24) clearly holds on Γ .

Choosing \mathbf{v} in (5.20) to be the basis function associated with node $(x_{i+1/2}, y_j)$, we have for $i = 1, \dots, \hat{N}_x - 1$,

$$\begin{aligned}\mathbf{Z}_{x,i+1/2j}^{*,n} &= \frac{1}{2}[K_{11}(p^n)_{i+1/2j+1/2} + K_{11}(p^n)_{i+1/2j-1/2}]\tilde{\mathbf{Z}}_{x,i+1/2j}^{*,n} \\ &\quad + \frac{1}{2}[\partial_t(K_{11}(p^n))_{i+1/2j+1/2} + \partial_t(K_{11}(p^n))_{i+1/2j-1/2}]\tilde{\mathbf{U}}_{x,i+1/2j}^{*,n}.\end{aligned}$$

Since $\partial_t \mathbf{u}^n = K(p^n)\partial_t \tilde{\mathbf{u}}^n + \partial_t K(p^n)\tilde{\mathbf{u}}^n$, Taylor's theorem gives for $i = 1, \dots, \hat{N}_x - 1$,

$$\begin{aligned}\partial_t \mathbf{u}_{x,i+1/2j}^n &= \frac{1}{2}[K_{11}(p^n)_{i+1/2j+1/2} + K_{11}(p^n)_{i+1/2j-1/2}]\partial_t \tilde{\mathbf{u}}_{x,i+1/2j}^n \\ &\quad + \frac{1}{2}[\partial_t(K_{11}(p^n))_{i+1/2j+1/2} + \partial_t(K_{11}(p^n))_{i+1/2j-1/2}]\tilde{\mathbf{u}}_{x,i+1/2j}^n + O(H^2).\end{aligned}$$

Therefore,

$$|\mathbf{Z}_{x,i+1/2j}^{*,n} - \partial_t \mathbf{u}_{x,i+1/2j}^n| \leq C|\tilde{\mathbf{Z}}_{x,i+1/2j}^{*,n} - \partial_t \tilde{\mathbf{u}}_{x,i+1/2j}^n| + O(H^2).$$

In a similar manner we can bound $|\mathbf{Z}_{y,ij+1/2}^{*,n} - \partial_t \mathbf{u}_{y,ij+1/2}^n|$, and (5.25) follows. \square

We can now extend a corollary from Arbogast, Wheeler and Yotov [8] to arrive at the following statement. For the $\tilde{\mathbf{U}}^{*,n}, P^{*,n}, \mathbf{Z}^{*,n}, \tilde{\mathbf{Z}}^{*,n}$ and $W^{*,n}$ in Lemma 5.3, there exists a constant C , independent of H , such that

$$\begin{aligned}\|\tilde{\mathbf{U}}^{*,n} - \tilde{\mathbf{u}}^n\|_{\text{TM}} &\leq CH^2, \\ \|\tilde{\mathbf{Z}}^{*,n} - \partial_t \tilde{\mathbf{u}}^n\|_{\text{TM}} &\leq CH^2, \\ \|\mathbf{Z}^{*,n} - \partial_t \mathbf{u}^n\|_{\text{TM}} &\leq CH^2.\end{aligned}$$

Lemma 5.4 There exists a constant C independent of H and Δt such that,

$$\|\nabla \cdot (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n)\| \leq CH, \quad (5.26)$$

$$\|d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n\|_{\text{TM}} + \|d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n\|_{\text{TM}} \leq C(H^2 + \Delta t). \quad (5.27)$$

Proof To prove this lemma, consider the time difference of (5.3)-(5.5),

$$(\nabla \cdot d_t \mathbf{u}^n, w) = (d_t b^n, w), \quad \forall w \in W, \quad (5.28)$$

$$(d_t \tilde{\mathbf{u}}^n, \mathbf{v}) = (d_t p^n, \nabla \cdot \mathbf{v}), \quad \forall \mathbf{v} \in V^0, \quad (5.29)$$

$$(d_t \mathbf{u}^n, \mathbf{v}) = (d_t(K(p^n))\tilde{\mathbf{u}}^n, \mathbf{v}) + (K(p^{n-1})d_t \tilde{\mathbf{u}}^n, \mathbf{v}), \quad \forall \mathbf{v} \in V, \quad (5.30)$$

and the time difference of (5.10)-(5.12),

$$(\nabla \cdot d_t \underline{\mathbf{U}}_H^n, w) = (d_t b^n, w), \quad \forall w \in W_H, \quad (5.31)$$

$$(d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v})_{\text{TM}} = (d_t \underline{P}_H^n, \nabla \cdot \mathbf{v}), \quad \forall \mathbf{v} \in V_H^0, \quad (5.32)$$

$$\begin{aligned} (d_t \underline{\mathbf{U}}_H^n, \mathbf{v})_{\text{TM}} &= (d_t(K(\mathcal{P}_H(p^n)))\tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v})_{\text{T}} \\ &\quad + (K(\mathcal{P}_H(p^{n-1}))d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v})_{\text{T}}, \quad \forall \mathbf{v} \in V_H. \end{aligned} \quad (5.33)$$

Subtract (5.31) from (5.28), and subtract (5.32) and (5.33) from (5.19) and (5.20) to give,

$$(\nabla \cdot (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n), w) = 0, \quad (5.34)$$

$$(\tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v})_{\text{TM}} = (W^{*,n} - d_t \underline{P}_H^n, \nabla \cdot \mathbf{v}), \quad (5.35)$$

$$\begin{aligned} (\mathbf{Z}^{*,n} - d_t \underline{\mathbf{U}}_H^n, \mathbf{v})_{\text{TM}} &= (K(p^n)\tilde{\mathbf{Z}}^{*,n} - K(\mathcal{P}_H(p^{n-1}))d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v})_{\text{T}} \\ &\quad + (\partial_t(K(p^n)\tilde{\mathbf{U}}^{*,n} - d_t K(\mathcal{P}_H(p^n))\tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v})_{\text{T}}. \end{aligned} \quad (5.36)$$

Using (5.34) and applying the Cauchy-Schwarz inequality we have,

$$\begin{aligned} \|\nabla \cdot (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n)\|^2 &= (\nabla \cdot (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n), \nabla \cdot (d_t \mathbf{u}^n - \Pi d_t \mathbf{u}^n)) \\ &\leq \|\nabla \cdot (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n)\| \|\nabla \cdot (d_t \mathbf{u}^n - \Pi d_t \mathbf{u}^n)\|. \end{aligned}$$

Thus, by (3.24) the first part of the lemma is obtained.

Now, let $\mathbf{v} = \Pi d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n$ in (5.35) and $\mathbf{v} = \tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n$ in (5.36), use (5.34) and combine to get,

$$\begin{aligned} & (K(p^n) \tilde{\mathbf{Z}}^{*,n} - K(\mathcal{P}_H(p^{n-1})) d_t \tilde{\underline{\mathbf{U}}}_H^n, \tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n)_T \\ &= -(\tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n, \Pi d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n)_{\text{TM}} + (\mathbf{Z}^{*,n} - d_t \underline{\mathbf{U}}_H^n, \tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n)_{\text{TM}} \\ & \quad - (\partial_t K(p^n) \tilde{\underline{\mathbf{U}}}^{*,n} - d_t K(\mathcal{P}_H(p^n)) \tilde{\underline{\mathbf{U}}}_H^n, \tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n)_T. \end{aligned} \quad (5.37)$$

Adding $(K(\mathcal{P}_H(p^{n-1})) \tilde{\mathbf{Z}}^{*,n}, \tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n)_T$ to both sides of (5.37), using the boundedness assumption on K , Taylor's Theorem, the Cauchy-Schwarz inequality and (4.2) results in,

$$\begin{aligned} \|\tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n\|_{\text{TM}} &\leq C(\|\Pi d_t \mathbf{u}^n - \mathbf{Z}^{*,n}\|_{\text{TM}} + \Delta t \|\tilde{\underline{\mathbf{U}}}^{*,n}\|_{\text{TM}} \\ & \quad + \|d_t(K(p^n) - K(\mathcal{P}_H(p^n))) \tilde{\underline{\mathbf{U}}}^{*,n}\|_T \\ & \quad + \|d_t K(\mathcal{P}_H(p^n))(\tilde{\underline{\mathbf{U}}}^{*,n} - \tilde{\underline{\mathbf{U}}}_H^n)\|_T + (H^2 + \Delta t) \|\tilde{\mathbf{Z}}^{*,n}\|_T). \end{aligned}$$

Taylor's theorem, the estimate (3.25) and Lemma 5.3 imply that $\|\Pi d_t \mathbf{u}^n - \mathbf{Z}^{*,n}\|_{\text{TM}} \leq C(H^2 + \Delta t)$. By Lemma 5.3 $\|\tilde{\underline{\mathbf{U}}}^{*,n}\|_{\text{TM}}$ and $\|\tilde{\mathbf{Z}}^{*,n}\|_T$ are bounded. Thus, by Taylor's theorem, the Lipschitz condition on $\partial_t K$, the boundedness of $\partial_t K$ and the approximation properties of \mathcal{P}_H ,

$$\|\tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n\|_{\text{TM}} \leq C(H^2 + \Delta t + \|\tilde{\underline{\mathbf{U}}}^{*,n} - \tilde{\underline{\mathbf{U}}}_H^n\|_T). \quad (5.38)$$

By results from Arbogast, Wheeler and Yotov [8], $\|\tilde{\underline{\mathbf{U}}}^{*,n} - \tilde{\underline{\mathbf{U}}}_H^n\|_T \leq CH^2$. Hence, by the triangle inequality and Lemma 5.3,

$$\|d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n\|_{\text{TM}} \leq C(H^2 + \Delta t).$$

Now, let $\mathbf{v} = \mathbf{Z}^{*,n} - d_t \underline{\mathbf{U}}_H^n$ in (5.36) and use the Cauchy-Schwarz inequality to get

$$\begin{aligned} \|\mathbf{Z}^{*,n} - d_t \underline{\mathbf{U}}_H^n\|_{\text{TM}} &\leq \|K(p^n) \tilde{\mathbf{Z}}^{*,n} - K(\mathcal{P}_H(p^{n-1})) d_t \tilde{\underline{\mathbf{U}}}_H^n\|_T \\ & \quad + \|\partial_t K(p^n) \tilde{\underline{\mathbf{U}}}^{*,n} - d_t(K(\mathcal{P}_H(p^n))) \tilde{\underline{\mathbf{U}}}_H^n\|_T. \end{aligned}$$

By the Lipschitz assumptions on K and $\partial_t K$, Taylor's theorem, the approximation properties of \mathcal{P}_H and the boundedness of $\partial_t K$,

$$\begin{aligned} & \|\mathbf{Z}^{*,n} - d_t \underline{\mathbf{U}}_H^n\|_{\text{TM}} \\ & \leq \|(K(p^n) - K(\mathcal{P}_H(p^{n-1}))) \tilde{\mathbf{Z}}^{*,n}\|_T + \|K(\mathcal{P}_H(p^{n-1}))(\tilde{\mathbf{Z}}^{*,n} - d_t \tilde{\underline{\mathbf{U}}}_H^n)\|_T \\ & \quad + \|(\partial_t K(p^n) - d_t K(\mathcal{P}_H(p^n))) \tilde{\underline{\mathbf{U}}}^{*,n}\|_T + \|d_t K(\mathcal{P}_H(p^n))(\tilde{\underline{\mathbf{U}}}^{*,n} - \tilde{\underline{\mathbf{U}}}_H^n)\|_T \\ & \leq C(H^2 + \Delta t). \end{aligned}$$

The triangle inequality and Lemma 5.3 result in,

$$\|d_t \dot{\mathbf{u}}^n - d_t \underline{\mathbf{U}}_H^n\|_{\text{TM}} \leq C(H^2 + \Delta t).$$

□

Remark 5.1.1 By the inverse assumption, the definition of Π and (3.25) we have

$$\begin{aligned} \|\tilde{\underline{\mathbf{U}}}_H^n\|_{L^\infty} &\leq \|\tilde{\underline{\mathbf{U}}}_H^n - \Pi \tilde{\mathbf{u}}^n\|_{L^\infty} + \|\Pi \tilde{\mathbf{u}}^n - \tilde{\mathbf{u}}^n\|_{L^\infty} + \|\tilde{\mathbf{u}}^n\|_{L^\infty} \\ &\leq CH^{-d/2} \|\tilde{\underline{\mathbf{U}}}_H^n - \Pi \tilde{\mathbf{u}}^n\|_{\text{TM}} + \|\Pi \tilde{\mathbf{u}}^n - \tilde{\mathbf{u}}^n\|_{L^\infty} + \|\tilde{\mathbf{u}}^n\|_{L^\infty} \\ &\leq C(H^{-d/2}H^2 + H + 1), \end{aligned}$$

where d is the space dimension. Thus, $\|\tilde{\underline{\mathbf{U}}}_H^n\|_{L^\infty}$ is bounded.

Proof (Of Theorem 5.2) Results (5.13) and (5.14) have been proven by Arbogast, Wheeler and Yotov [8].

In order to derive (5.15), subtract (5.32)-(5.33) from (5.29)-(5.30) and use the definition of the L^2 projection to give,

$$(d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v}) + E_{\text{TM}}(d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v}) = (d_t \hat{p}_H^n - d_t \underline{P}_H^n, \nabla \cdot \mathbf{v}), \mathbf{v} \in V_H^0, \quad (5.39)$$

$$\begin{aligned} (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n, \mathbf{v}) + E_{\text{TM}}(d_t \underline{\mathbf{U}}_H^n, \mathbf{v}) &= (d_t(K(p^n))\tilde{\mathbf{u}}^n - d_t(K(\mathcal{P}_H(p^n)))\tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v}) \\ &\quad + (K(p^{n-1})d_t \tilde{\mathbf{u}}^n - K(\mathcal{P}_H(p^{n-1}))d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v}) \\ &\quad + E_{\text{T}}(d_t(K(\mathcal{P}_H(p^n)))\tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v}) \\ &\quad + E_{\text{T}}(K(\mathcal{P}_H(p^{n-1}))d_t \tilde{\underline{\mathbf{U}}}_H^n, \mathbf{v}), \mathbf{v} \in V_H. \end{aligned} \quad (5.40)$$

Let ϕ satisfy the auxilliary problem with $\rho^n \in L^2(\Omega)$ defined later,

$$-\nabla \cdot K(\mathcal{P}_H(p^{n-1}))\nabla \phi^n = \rho^n, \Omega, \quad (5.41)$$

$$K(\mathcal{P}_H(p^{n-1}))\nabla \phi^n \cdot \mathbf{n} = 0, \Gamma, \quad (5.42)$$

where we assume that $\int_\Omega \rho^n = 0$. Elliptic regularity implies that

$$\|\phi^n\|_2 \leq C\|\rho^n\|. \quad (5.43)$$

By equations (5.41) and (5.39) and the definition of Π ,

$$(d_t \hat{p}_H^n - d_t \underline{P}_H^n, \rho^n)$$

$$\begin{aligned}
&= -(d_t \hat{p}_H^n - d_t \underline{P}_H^n, \nabla \cdot \Pi K(\mathcal{P}_H(p^{n-1})) \nabla \phi^n) \\
&= -(d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n, \Pi K(\mathcal{P}_H(p^{n-1})) \nabla \phi^n) \\
&\quad - E_{\text{TM}}(d_t \tilde{\underline{\mathbf{U}}}_H^n, \Pi K(\mathcal{P}_H(p^{n-1})) \nabla \phi^n) \\
&= -(d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n, \Pi K(\mathcal{P}_H(p^{n-1})) \nabla \phi^n - K(\mathcal{P}_H(p^{n-1})) \nabla \phi^n) \\
&\quad - (K(\mathcal{P}_H(p^{n-1}))(d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n), \nabla \phi^n - \Pi \nabla \phi^n) \\
&\quad - (K(\mathcal{P}_H(p^{n-1}))(d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n), \Pi \nabla \phi^n) \\
&\quad - E_{\text{TM}}(d_t \tilde{\underline{\mathbf{U}}}_H^n, \Pi K(\mathcal{P}_H(p^{n-1})) \nabla \phi^n). \tag{5.44}
\end{aligned}$$

By (5.40),

$$\begin{aligned}
&-(K(\mathcal{P}_H(p^{n-1}))(d_t \tilde{\mathbf{u}}^n - d_t \tilde{\underline{\mathbf{U}}}_H^n), \Pi \nabla \phi^n) \\
&= ((K(p^{n-1}) - K(\mathcal{P}_H(p^{n-1})))d_t \tilde{\mathbf{u}}^n, \Pi \nabla \phi^n) - (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n, \Pi \nabla \phi^n) \\
&\quad + (d_t(K(p^n))\tilde{\mathbf{u}}^n - d_t(K(\mathcal{P}_H(p^n)))\tilde{\underline{\mathbf{U}}}_H^n, \Pi \nabla \phi^n) - E_{\text{TM}}(d_t \underline{\mathbf{U}}_H^n, \Pi \nabla \phi^n) \\
&\quad + E_{\text{T}}(d_t(K(\mathcal{P}_H(p^n)))\tilde{\underline{\mathbf{U}}}_H^n, \Pi \nabla \phi^n) \\
&\quad + E_{\text{T}}(K(\mathcal{P}_H(p^{n-1}))d_t \tilde{\underline{\mathbf{U}}}_H^n, \Pi \nabla \phi^n). \tag{5.45}
\end{aligned}$$

We also have by integration by parts, (5.34) and (5.36)

$$\begin{aligned}
&-(d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n, \Pi \nabla \phi^n) \\
&= -(d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n, \Pi \nabla \phi^n - \nabla \phi^n) - (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n, \nabla \phi^n) \\
&= -(d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n, \Pi \nabla \phi^n - \nabla \phi^n) + (\nabla \cdot (d_t \mathbf{u}^n - d_t \underline{\mathbf{U}}_H^n), \phi^n - \hat{\phi}_H^n). \tag{5.46}
\end{aligned}$$

Furthermore, we can write

$$\begin{aligned}
&(d_t K(p^n)\tilde{\mathbf{u}}^n - d_t K(\mathcal{P}_H(p^n))\tilde{\underline{\mathbf{U}}}_H^n, \Pi \nabla \phi^n) \\
&= ((d_t K(p^n) - d_t K(\mathcal{P}_H(p^n)))\tilde{\mathbf{u}}^n, \Pi \nabla \phi^n) \\
&\quad + (d_t K(\mathcal{P}_H(p^n))(\tilde{\mathbf{u}}^n - \tilde{\underline{\mathbf{U}}}_H^n), \Pi \nabla \phi^n). \tag{5.47}
\end{aligned}$$

Using (3.16) and recalling the definition of the lowest order RTN space gives,

$$\begin{aligned}
|E_{\text{TM}}(d_t \tilde{\underline{\mathbf{U}}}_H^n, \Pi K(p^n) \nabla \phi^n)| &\leq C \sum_{E \in \mathcal{T}_H} \sum_{|\alpha|=2} \left\| \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} (d_t \tilde{\underline{\mathbf{U}}}_H^n \cdot \Pi K(p^n) \nabla \phi^n) \right\|_{L^1(E)} H^2 \\
&\leq C \sum_{E \in \mathcal{T}_H} \left(\left\| \frac{\partial d_t \tilde{\underline{\mathbf{U}}}_{H,x}^n}{\partial x} \right\|_{0,E} \left\| \frac{\partial}{\partial x} (\Pi K(p^n) \nabla \phi^n)^x \right\|_{0,E} \right. \\
&\quad \left. + \left\| \frac{\partial d_t \tilde{\underline{\mathbf{U}}}_{H,y}^n}{\partial y} \right\|_{0,E} \left\| \frac{\partial}{\partial y} (\Pi K(p^n) \nabla \phi^n)^y \right\|_{0,E} \right) H^2. \tag{5.48}
\end{aligned}$$

By Lemma 5.2 and the inverse inequality we have

$$\begin{aligned}
\left\| \frac{\partial d_t \tilde{\mathbf{U}}_{H,x}^n}{\partial x} \right\|_{0,E} &\leq \left\| \frac{\partial}{\partial x} (d_t \tilde{\mathbf{U}}_{H,x}^n - \Pi d_t \tilde{\mathbf{u}}_x^n) \right\|_{0,E} + \left\| \frac{\partial}{\partial x} \Pi d_t \tilde{\mathbf{u}}_x^n \right\|_{0,E} \\
&\leq C \|d_t \tilde{\mathbf{U}}_{H,x}^n - \Pi d_t \tilde{\mathbf{u}}_x^n\|_{0,E} H^{-1} + \left\| \frac{\partial}{\partial x} d_t \tilde{\mathbf{u}}_x^n \right\|_{0,E} \\
&\leq C \|d_t \tilde{\mathbf{U}}_H^n - \Pi d_t \tilde{\mathbf{u}}^n\|_{0,E} H^{-1} + \|d_t \tilde{\mathbf{u}}^n\|_{1,E}.
\end{aligned} \tag{5.49}$$

Since $d_t \tilde{\mathbf{U}}_H^n$ and $\Pi K(p^n) \nabla \phi^n$ are in V_H , by (5.48) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
|E_{\text{TM}}(d_t \tilde{\mathbf{U}}_H^n, \Pi K(p^n) \nabla \phi^n)| &\leq C \sum_E (\|d_t \tilde{\mathbf{U}}_H^n - \Pi d_t \tilde{\mathbf{u}}^n\|_{0,E} + \|d_t \tilde{\mathbf{u}}^n\|_{1,E} H) \|\phi^n\|_{2,E} H \\
&\leq C (\|d_t \tilde{\mathbf{U}}_H^n - \Pi d_t \tilde{\mathbf{u}}^n\|_0 + \|d_t \tilde{\mathbf{u}}^n\|_1 H) \|\phi^n\|_2 H.
\end{aligned} \tag{5.50}$$

As done above and noting that K has bounded second derivatives,

$$\begin{aligned}
|E_{\text{TM}}(d_t \mathbf{U}_H^n, \Pi \nabla \phi^n)| &\leq C (\|d_t \mathbf{U}_H^n - \Pi d_t \mathbf{u}^n\| + \|d_t \mathbf{u}^n\|_1 H) \\
&\quad \times \|\phi^n\|_2 H,
\end{aligned} \tag{5.51}$$

$$\begin{aligned}
|E_{\text{T}}(d_t(K(\mathcal{P}_H(p^n))) \tilde{\mathbf{U}}_H^n, \Pi \nabla \phi^n)| &\leq (C + \Delta t) (\|\tilde{\mathbf{U}}_H^n - \Pi \tilde{\mathbf{u}}^n\| + \|\tilde{\mathbf{u}}^n\|_1 H) \\
&\quad \times \|\phi^n\|_2 H,
\end{aligned} \tag{5.52}$$

$$\begin{aligned}
|E_{\text{T}}(K(\mathcal{P}_H(p^{n-1})) d_t \tilde{\mathbf{U}}_H^n, \Pi \nabla \phi^n)| &\leq C (\|d_t \tilde{\mathbf{U}}_H^n - \Pi d_t \tilde{\mathbf{u}}^n\| + \|d_t \tilde{\mathbf{u}}^n\|_1 H) \\
&\quad \times \|\phi^n\|_2 H,
\end{aligned} \tag{5.53}$$

where C in the second and third inequalities depends on K^* .

Combining (5.44) with (5.45)-(5.53), applying approximation properties of the L^2 and Π projections, using Lemma 5.4, and equations (5.13) and (3.25) gives

$$\begin{aligned}
&(d_t \hat{p}_H^n - d_t \underline{p}_H^n, \rho^n) \\
&\leq \|d_t \tilde{\mathbf{u}}^n - d_t \tilde{\mathbf{U}}_H^n\| \|\Pi K(p^n) \nabla \phi^n - K(p^n) \nabla \phi^n\| \\
&\quad + \|K(\mathcal{P}_H(p^{n-1}))(d_t \tilde{\mathbf{u}}^n - d_t \tilde{\mathbf{U}}_H^n)\| \|\nabla \phi^n - \Pi \nabla \phi^n\| \\
&\quad + \|(d_t K(p^n) - d_t K(\mathcal{P}_H(p^n))) \tilde{\mathbf{u}}^n\| \|\Pi \nabla \phi^n\| \\
&\quad + \|d_t K(\mathcal{P}_H(p^n))(\tilde{\mathbf{u}}^n - \tilde{\mathbf{U}}_H^n)\| \|\Pi \nabla \phi^n\| \\
&\quad + \|(K(p^{n-1}) - K(\mathcal{P}_H(p^{n-1}))) d_t \tilde{\mathbf{u}}^n\| \|\Pi \nabla \phi^n\| \\
&\quad + \|d_t \mathbf{u}^n - d_t \mathbf{U}_H^n\| \|\Pi \nabla \phi^n - \nabla \phi^n\| \\
&\quad + \|\nabla \cdot (d_t \mathbf{u}^n - d_t \mathbf{U}_H^n)\| \|\phi^n - \hat{\phi}_H^n\| \\
&\quad + \|d_t(K(p^n)) \tilde{\mathbf{u}}^n - d_t(K(\mathcal{P}_H(p^n))) \tilde{\mathbf{U}}_H^n\| \|\Pi \nabla \phi^n\|
\end{aligned}$$

$$\begin{aligned}
& + |E_{\text{TM}}(d_t \tilde{\mathbf{U}}_H^n, \Pi K(p^n) \nabla \phi^n)| + |E_{\text{TM}}(d_t \mathbf{U}_H^n, \Pi \nabla \phi^n)| \\
& + |E_{\text{T}}(d_t(K(\mathcal{P}_H(p^n))) \tilde{\mathbf{U}}_H^n, \Pi \nabla \phi^n)| + |E_{\text{T}}(K(\mathcal{P}_H(p^{n-1})) d_t \tilde{\mathbf{U}}_H^n, \Pi \nabla \phi^n)| \\
& \leq C(H^2 + \Delta t) \|\phi^n\|_2.
\end{aligned}$$

Taking $\rho^n = d_t \hat{p}^n - d_t \underline{P}_H^n$ and noting that $\int_{\Omega} \rho^n = \int_{\Omega} \rho^0 = 0$ and applying equation (5.43) we have,

$$\|d_t \hat{p}_H^n - d_t \underline{P}_H^n\| \leq C(H^2 + \Delta t). \quad (5.54)$$

□

Convergence Estimate of the Nonlinear Scheme

We now prove the following theorem about the convergence of the above coarse grid finite difference scheme:

Theorem 5.3 Let $P_H^n, \tilde{\mathbf{U}}_H^n$ and $\mathbf{U}_H^n, n = 1, \dots, N$ be defined as in (5.6)-(5.8) with initial values $P_H^0 = \hat{p}_H(t^0, \cdot)$. Then, there exists a positive constant C , independent of H and Δt such that

$$\|P_H^N - p^N\|_{\text{M}} + \{\Delta t \sum_{n=1}^N K_* \|\tilde{\mathbf{U}}_H^n - \tilde{u}^n\|_{\text{T}}^2\}^{1/2} \leq C(H^2 + \Delta t). \quad (5.55)$$

Proof Let $\gamma^n = P_H^n - \underline{P}_H^n, \boldsymbol{\eta}^n = \tilde{\mathbf{U}}_H^n - \tilde{\mathbf{U}}_H^n, \boldsymbol{\xi}^n = \mathbf{U}_H^n - \underline{\mathbf{U}}_H^n$ and $\alpha^n = \underline{P}_H^n - p^n$. Subtracting $(d_t \underline{P}_H^n, w) + (\nabla \cdot \underline{\mathbf{U}}_H^n, w)$ from both sides of (5.6) and using equations (5.10) and (5.3) we have,

$$\begin{aligned}
(d_t \gamma^n, w) + (\nabla \cdot \boldsymbol{\xi}^n, w) &= (f^n, w) - (\nabla \cdot \underline{\mathbf{U}}_H^n, w) - (d_t \underline{P}_H^n, w) \\
&= (\partial_t p^n, w) - (d_t p^n, w) - (d_t \alpha^n, w) \\
&= (\epsilon^n, w) - (d_t \alpha^n, w), \quad w \in W_H,
\end{aligned} \quad (5.56)$$

where ϵ^n is a time truncation term. Subtracting (5.11) from (5.7) results in,

$$(\boldsymbol{\eta}^n, \mathbf{v})_{\text{TM}} = (\gamma^n, \nabla \cdot \mathbf{v}), \quad \mathbf{v} \in V_H^0, \quad (5.57)$$

and subtracting (5.12) from (5.8) gives,

$$\begin{aligned}
(\boldsymbol{\xi}^n, \mathbf{v})_{\text{TM}} &= (K(\mathcal{P}_H(P_H^n)) \tilde{\mathbf{U}}_H^n, \mathbf{v})_{\text{T}} - (K(\mathcal{P}_H(p^n)) \tilde{\mathbf{U}}_H^n, \mathbf{v})_{\text{T}} \\
&= (K(\mathcal{P}_H(P_H^n)) \boldsymbol{\eta}^n, \mathbf{v})_{\text{T}} + ((K(\mathcal{P}_H(P_H^n)) - K(\mathcal{P}_H(\underline{P}_H^n))) \tilde{\mathbf{U}}_H^n, \mathbf{v})_{\text{T}} \\
&\quad - ((K(\mathcal{P}_H(p^n)) - K(\mathcal{P}_H(\underline{P}_H^n))) \tilde{\mathbf{U}}_H^n, \mathbf{v})_{\text{T}}, \quad \mathbf{v} \in V_H.
\end{aligned} \quad (5.58)$$

Letting $w = \gamma^n$ in (5.56), $\mathbf{v} = \boldsymbol{\xi}^n$ in (5.57) and $\mathbf{v} = \boldsymbol{\eta}^n$ in (5.58) gives,

$$(d_t \gamma^n, \gamma^n) = -(\nabla \cdot \boldsymbol{\xi}^n, \gamma^n) + (\epsilon^n, \gamma^n)_M - (d_t \alpha^n, \gamma^n), \quad (5.59)$$

$$(\boldsymbol{\eta}^n, \boldsymbol{\xi}^n)_{\text{TM}} = (\gamma^n, \nabla \cdot \boldsymbol{\xi}^n), \quad (5.60)$$

$$\begin{aligned} (\boldsymbol{\xi}^n, \boldsymbol{\eta}^n)_{\text{TM}} &= (K(\mathcal{P}_H(P_H^n))\boldsymbol{\eta}^n, \boldsymbol{\eta}^n)_T + ((K(\mathcal{P}_H(P_H^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n, \boldsymbol{\eta}^n)_T \\ &\quad - ((K(\mathcal{P}_H(p^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n, \boldsymbol{\eta}^n)_T. \end{aligned} \quad (5.61)$$

Combining equations (5.59)-(5.61), applying the Cauchy-Schwarz inequality and (4.2) we have,

$$\begin{aligned} &\frac{1}{2\Delta t} [\|\gamma^n\|_M^2 - \|\gamma^{n-1}\|_M^2] + \|K(\mathcal{P}_H(P_H^n))^{1/2}\boldsymbol{\eta}^n\|_T^2 \\ &\leq (d_t \gamma^n, \gamma^n)_M + \|K(\mathcal{P}_H(P_H^n))^{1/2}\boldsymbol{\eta}^n\|_T^2 \\ &\leq (\epsilon^n, \gamma^n)_M - (d_t \alpha^n, \gamma^n)_M + ((K(\mathcal{P}_H(p^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n, \boldsymbol{\eta}^n)_T \\ &\quad - ((K(\mathcal{P}_H(P_H^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n, \boldsymbol{\eta}^n)_T \\ &\leq \frac{1}{2}\|\epsilon^n\|_M^2 + \|\gamma^n\|_M^2 + \frac{1}{2}\|d_t \alpha^n\|_M^2 + C\|(K(\mathcal{P}_H(p^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n\|_T^2 \\ &\quad + C\|(K(\mathcal{P}_H(P_H^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n\|_T^2 + \delta\|\boldsymbol{\eta}^n\|_{\text{TM}}^2, \end{aligned}$$

where $\delta \leq K_*/2$.

Now,

$$|\epsilon_{ij}^n| = \frac{1}{\Delta t} \left| \int_{t^{n-1}}^{t^n} p_{it}(x_i, y_j, t)(t - t^n) dt \right| \leq \|p_{it}(x_i, y_j, \cdot)\|_{L^2(t^{n-1}, t^n)} (\Delta t)^{\frac{1}{2}}.$$

So,

$$\|\epsilon^n\|_M^2 \leq \Delta t \sum_{ij} H_i^x H_j^y \|p_{it}(x_i, y_j, \cdot)\|_{L^2(t^{n-1}, t^n)}^2. \quad (5.62)$$

By the triangle inequality and Theorem 5.2,

$$\|d_t \alpha^n\|_M^2 \leq C(H^4 + \Delta t^2).$$

By the Lipschitz assumption on K , the definition of \mathcal{P}_H and Theorem 5.2

$$\|(K(\mathcal{P}_H(p^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n\|_T^2 \leq C\|p^n - \underline{P}_H^n\|_M^2 \leq CH^4, \quad (5.63)$$

$$\|(K(\mathcal{P}_H(P_H^n)) - K(\mathcal{P}_H(\underline{P}_H^n)))\tilde{\mathbf{U}}_H^n\|_T^2 \leq C\|\gamma^n\|_M^2, \quad (5.64)$$

where we have used the boundedness of $\|\tilde{\mathbf{U}}_H^n\|_{L^\infty}$ as per Remark 5.1.1.

Multiplying by $2\Delta t$, bringing the $\delta\|\boldsymbol{\eta}^n\|_{\text{TM}}^2$ term to the left-hand side, summing on $n, n = 1, \dots, N$, using (5.62)-(5.64) and applying Gronwall's Lemma 3.1 gives,

$$\begin{aligned} & \|\gamma^N\|_{\text{M}}^2 - \|\gamma^0\|_{\text{M}}^2 + \Delta t \sum_{n=1}^N \|K(\mathcal{P}_H(P_H^n))^{1/2} \boldsymbol{\eta}^n\|_{\text{T}}^2 \\ & \leq C\Delta t \sum_{n=1}^N (\|\epsilon^n\|_{\text{M}}^2 + \|d_t \alpha^n\|_{\text{M}}^2 + \|\alpha^n\|_{\text{M}}^2) + CH^4 + C\Delta t \sum_{n=1}^N \|\gamma^n\|_{\text{M}}^2 \\ & \leq C(\Delta t^2 + H^4). \end{aligned}$$

The proof is completed by applying the initial conditions on P_H^0 and \underline{P}_H^0 , Theorem 5.2 and the triangle inequality. \square

5.1.2 Fine Grid Linear Scheme

We now consider a linear cell-centered finite difference scheme on the fine grid where we make use of the nonlinear solution on the coarse grid. Note that since we assume \mathcal{T}_h is a refinement of \mathcal{T}_H , we have, $W_H \subset W_h$ and $\mathbf{V}_H \subset \mathbf{V}_h$.

We solve the following problem for $P_h^n \in W_h$, $\tilde{\mathbf{U}}_h^n \in V_h$ and $\mathbf{U}_h^n \in V_H^0$ at each time step $n = 1, \dots, N$,

$$(d_t P_h^n, w) = -(\nabla \cdot \mathbf{U}_h^n, w) + (f^n, w), w \in W_h, \quad (5.65)$$

$$(\tilde{\mathbf{U}}_h^n, \mathbf{v})_{\text{TM}} = (P_h^n, \nabla \cdot \mathbf{v}), \mathbf{v} \in V_h^0, \quad (5.66)$$

$$\begin{aligned} (\mathbf{U}_h^n, \mathbf{v})_{\text{TM}} &= (K(\mathcal{P}_H(P_H^n)) \tilde{\mathbf{U}}_h^n, \mathbf{v})_{\text{T}} \\ &+ (K'(\mathcal{P}_H(P_H^n)) \mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) (\mathcal{P}_h(P_h^n) - \mathcal{P}_H(P_H^n)), \mathbf{v})_{\text{T}}, \mathbf{v} \in V_h. \end{aligned} \quad (5.67)$$

We define $\mathcal{Q}_H(\tilde{\mathbf{u}})$ as a vector quantity with entries $\mathcal{Q}_H^x(\tilde{u}^x)$ and $\mathcal{Q}_H^y(\tilde{u}^y)$. The entry $\mathcal{Q}_H^x(\tilde{u}^x)$ is defined from the values of $\tilde{u}_{i+1/2,j}^x$ for $i = 0, \dots, \hat{N}_x$ and $j = 1, \dots, \hat{N}_y$ as follows. For points (x, y) such that $x_{i-1/2} \leq x \leq x_{i+1/2}, i \in \{1, \dots, \hat{N}_x\}$ and $y_j \leq y \leq y_{j+1}, j \in \{1, \dots, \hat{N}_y\}$, we take $\mathcal{Q}_H^x(\tilde{u}^x)$ to be the bilinear interpolant of $\tilde{u}_{i-1/2,j}^x, \tilde{u}_{i+1/2,j}^x, \tilde{u}_{i-1/2,j+1}^x$ and $\tilde{u}_{i+1/2,j+1}^x$. This leaves a strip half a cell in height along the top and bottom of the domain. We will consider the bottom strip. For $i = 0, \dots, \hat{N}_x$, we set

$$\mathcal{Q}_H^x(\tilde{u}^x)(x_{i+1/2}, y_{1/2}) = \frac{(2H_1^y + H_2^y)\tilde{u}_{i+1/2,1}^x - H_1^y \tilde{u}_{i+1/2,2}^x}{H_1^y + H_2^y}.$$

Now, for points (x, y) such that $x_{i-1/2} \leq x \leq x_{i+1/2}, i \in \{1, \dots, \hat{N}_x\}$ and $y_{1/2} \leq y \leq y_1$, we let $\mathcal{Q}_H^x(\tilde{u}^x)(x, y)$ be the bilinear interpolant of the two interpolated values

$\mathcal{Q}_H^x(\tilde{u}^x)(x_{i-1/2}, y_{1/2})$ and $\mathcal{Q}_H^x(\tilde{u}^x)(x_{i+1/2}, y_{1/2})$ and the two values $\tilde{u}_{i-1/2,1}^x$ and $\tilde{u}_{i+1/2,1}^x$. An analogous definition is made along the top strip of the domain. The definition of $\mathcal{Q}_H^y(\tilde{u}^y)$ is similar to the above, except that the strips are along the left and right sides of the domain.

The following lemma summarizes the approximation error of \mathcal{Q}_H .

Lemma 5.5 If each component of $\tilde{\mathbf{u}}$ is twice differentiable, then for $\mathcal{Q}_H(\tilde{\mathbf{u}})$ defined above,

$$\|\mathcal{Q}_H(\tilde{\mathbf{u}}) - \tilde{\mathbf{u}}\|_{L^\infty} \leq CH^2.$$

Proof By Taylor's theorem we have that the two point extrapolation for the boundary points described above is $O(H^2)$ accurate. Thus, since bilinear interpolation is also $O(H^2)$ accurate, the lemma is proven. \square

We turn now to an analysis of the fine grid scheme.

Theorem 5.4 Let $P_h^n, \tilde{\mathbf{U}}_h^n$ and $\mathbf{U}_h^n, n = 1, \dots, N$ be defined as in (5.65)-(5.67) with initial values $P_h^0 = \hat{p}_h(t^0, \cdot)$. Then, there exists a positive constant C , independent of h, H and Δt such that

$$\begin{aligned} \|P_h^N - p^N\|_M + \left\{ \Delta t \sum_{n=1}^N K_* \|\tilde{\mathbf{U}}_h^n - \tilde{\mathbf{u}}^n\|_T^2 \right\}^{1/2} \\ \leq C(H^{4-d/2} + h^2 + \Delta t). \end{aligned}$$

Proof We can define $\underline{P}_h^n \in W_h, \tilde{\mathbf{U}}_h^n \in V_h$ and $\underline{\mathbf{U}}_h^n \in V_h^n$ at each $n = 1, \dots, N$ satisfying equations (5.10)-(5.12) and Theorem 5.2 on the fine grid.

Let $\gamma^n = P_h^n - \underline{P}_h^n, \boldsymbol{\eta}^n = \tilde{\mathbf{U}}_h^n - \tilde{\mathbf{U}}_h^n, \boldsymbol{\xi}^n = \mathbf{U}_h^n - \underline{\mathbf{U}}_h^n$ and $\alpha^n = \underline{P}_h^n - p^n$. As done in Theorem 5.3, we subtract $(d_t \underline{P}_h^n, w) + (\nabla \cdot \underline{\mathbf{U}}_h^n, w)$ from both sides of equation (5.65) and combine with equation (5.10) applied to the fine grid. We also subtract (5.11) and (5.12) from (5.66) and (5.67) to give the error equations,

$$(d_t \gamma^n, w) = -(\nabla \cdot \boldsymbol{\xi}^n, w) + (\epsilon^n, w) - (d_t \alpha^n, w), \quad (5.68)$$

$$(\boldsymbol{\eta}^n, \mathbf{v})_{\text{TM}} = (\gamma^n, \nabla \cdot \mathbf{v}), \quad (5.69)$$

$$\begin{aligned} (\boldsymbol{\xi}^n, \mathbf{v})_{\text{TM}} &= (K(\mathcal{P}_H(P_H^n))\tilde{\mathbf{U}}_h^n, \mathbf{v})_T - (K(\mathcal{P}_h(p^n))\tilde{\mathbf{U}}_h^n, \mathbf{v})_T \\ &\quad + (K'(\mathcal{P}_H(P_H^n))\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n)(\mathcal{P}_h(P_h^n) - \mathcal{P}_H(P_H^n)), \mathbf{v})_T. \end{aligned} \quad (5.70)$$

Using Taylor's Theorem, $K(\mathcal{P}_h(p^n))$ can be written as

$$K(\mathcal{P}_h(p^n)) = K(\mathcal{P}_h(P_H^n)) + K'(\mathcal{P}_H(P_H^n))(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n)) \\ + \frac{K''(\theta^n)}{2}(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n))^2,$$

where θ^n is between $\mathcal{P}_h(p^n)$ and $\mathcal{P}_H(P_H^n)$.

Using this expression in (5.70) and adding and subtracting the derivative term, $(K'(\mathcal{P}_H(P_H^n))\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n)\mathcal{P}_h(p^n), \mathbf{v})_{\mathbf{T}}$ gives,

$$(\boldsymbol{\xi}^n, \mathbf{v})_{\mathbf{T}\mathbf{M}} = (K(\mathcal{P}_H(P_H^n))\boldsymbol{\eta}^n, \mathbf{v})_{\mathbf{T}} \\ + (K'(\mathcal{P}_H(P_H^n))(\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \tilde{\mathbf{U}}_h^n)(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n)), \mathbf{v})_{\mathbf{T}} \\ + (K'(\mathcal{P}_H(P_H^n))\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n)(\mathcal{P}_h(P_h^n) - \mathcal{P}_h(p^n)), \mathbf{v})_{\mathbf{T}} \\ + (\frac{K''(\theta^n)}{2}(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n))^2\tilde{\mathbf{U}}_h^n, \mathbf{v})_{\mathbf{T}}.$$

Let $w = \gamma^n$, $\mathbf{v} = \boldsymbol{\xi}^n$ and $\mathbf{v} = \boldsymbol{\eta}^n$ in (5.68), (5.69) and (5.71), respectively, and combine to give,

$$\frac{1}{2\Delta t}[\|\gamma^n\|^2 - \|\gamma^{n-1}\|^2] + K_*\|\boldsymbol{\eta}^n\|_{\mathbf{T}}^2 \\ \leq (d_t\gamma^n, \gamma^n) + \|K(\mathcal{P}_H(P_H^n))^{1/2}\boldsymbol{\eta}^n\|_{\mathbf{T}}^2 \\ \leq \frac{1}{2}\|\epsilon^n\|^2 + \|\gamma^n\|^2 + \frac{1}{2}\|d_t\alpha^n\|^2 + \delta\|\boldsymbol{\eta}^n\|^2 \\ + C\|K'(\mathcal{P}_H(P_H^n))(\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \tilde{\mathbf{U}}_h^n)(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n))\|_{\mathbf{T}}^2 \\ + C\|K'(\mathcal{P}_H(P_H^n))\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n)(\mathcal{P}_h(P_h^n) - \mathcal{P}_h(p^n))\|_{\mathbf{T}}^2 \\ + C\|\frac{K''(\theta^n)}{2}(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n))^2\tilde{\mathbf{U}}_h^n\|_{\mathbf{T}}^2, \quad (5.71)$$

where $\delta \leq K_*/2$.

Consider now the last three terms of (5.71). The first of these can be bounded as follows.

$$\|K'(\mathcal{P}_H(P_H^n))(\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \tilde{\mathbf{U}}_h^n)(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n))\|_{\mathbf{T}}^2 \\ \leq C\|\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \tilde{\mathbf{U}}_h^n\|_{\mathbf{T}\mathbf{M}}^2\|\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n)\|_{L^\infty}^2, \quad (5.72)$$

where,

$$\|\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \tilde{\mathbf{U}}_h^n\|_{\mathbf{T}\mathbf{M}}^2 \leq \|\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \mathcal{Q}_H(\tilde{\mathbf{u}}^n)\|_{\mathbf{T}\mathbf{M}}^2 + \|\mathcal{Q}_H(\tilde{\mathbf{u}}^n) - \tilde{\mathbf{u}}^n\|_{\mathbf{T}\mathbf{M}}^2 \\ + \|\tilde{\mathbf{u}}^n - \tilde{\mathbf{U}}_h^n\|_{\mathbf{T}\mathbf{M}}^2.$$

Since $\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n)$ is a bilinear interpolant of terms that can be expressed in terms of nodal values of $\tilde{\mathbf{U}}_H^n$ on the coarse grid, it can be shown that,

$$\|\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \mathcal{Q}_H(\tilde{\mathbf{u}}^n)\|_{\text{TM}}^2 \leq C \|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{\text{TM},H}^2,$$

where $\|\cdot\|_{\text{TM},H}$ denotes the midpoint by trapezoidal norm on the coarse grid. Also, $\|\mathcal{Q}_H(\tilde{\mathbf{u}}^n) - \tilde{\mathbf{u}}^n\|_{\text{TM}}^2 \leq CH^4$ by Lemma 5.5. In order to bound the second term in (5.72), write it as,

$$\begin{aligned} \|\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n)\|_{L^\infty}^2 &\leq \|\mathcal{P}_H(P_H^n) - \mathcal{P}_H(p^n)\|_{L^\infty}^2 + \|\mathcal{P}_H(p^n) - p^n\|_{L^\infty}^2 \\ &\quad + \|p^n - \mathcal{P}_h(p^n)\|_{L^\infty}^2. \end{aligned}$$

By the definition of \mathcal{P}_H , the quasi-uniformity assumption on the coarse grid and Theorem 5.3,

$$\begin{aligned} \|\mathcal{P}_H(P_H^n) - \mathcal{P}_H(p^n)\|_{L^\infty}^2 &\leq \frac{C}{H^d} \|P_H^n - p^n\|_{M,H}^2 \\ &\leq CH^{-d}(H^4 + \Delta t^2), \end{aligned}$$

where d is the space dimension. By Lemma 5.1, $\|\mathcal{P}_H(p^n) - p^n\|_{L^\infty}^2 \leq CH^4$ and $\|\mathcal{P}_h(p^n) - p^n\|_{L^\infty}^2 \leq Ch^4$. Thus,

$$\begin{aligned} &\|K'(\mathcal{P}_H(P_H^n))(\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n) - \tilde{\mathbf{u}}_h^n)(\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n))\|_{\text{T}}^2 \\ &\leq C(\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{\text{TM},H}^2 + \|\tilde{\mathbf{u}}^n - \tilde{\mathbf{u}}_h^n\|_{\text{TM}}^2 + H^4)(H^{4-d} + H^{-d}\Delta t^2 + h^4). \end{aligned} \quad (5.73)$$

The second to last term in (5.71) can be bounded by,

$$\begin{aligned} &\|K'(\mathcal{P}_H(P_H^n))\mathcal{Q}_H(\tilde{\mathbf{U}}_H^n)(\mathcal{P}_h(P_h^n) - \mathcal{P}_h(p^n))\|_{\text{T}}^2 \\ &\leq C\|\tilde{\mathbf{U}}_H^n\|_{L^\infty}^2\|\mathcal{P}_h(P_h^n) - \mathcal{P}_h(p^n)\|_{\text{T}}^2 \\ &\leq C(H^{-d}\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{\text{TM}}^2 + \|\tilde{\mathbf{u}}^n\|_{L^\infty}^2)(\|P_h^n - \underline{P}_h^n\|_{\text{M}}^2 + \|\underline{P}_h^n - p^n\|_{\text{M}}^2) \\ &\leq C(H^{-d}\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{\text{TM}}^2 + C)(\|\gamma^n\|^2 + h^4). \end{aligned} \quad (5.74)$$

The last term in (5.71) is bounded by,

$$\begin{aligned} &\left\| \frac{K''(\theta^n)}{2} (\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n))^2 \tilde{\mathbf{u}}_h^n \right\|_{\text{T}}^2 \\ &\leq C\|\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n)\|_{L^\infty}^2\|\mathcal{P}_h(p^n) - \mathcal{P}_H(P_H^n)\|_{\text{T}}^2 \\ &\leq C(H^{4-d} + H^{-d}\Delta t^2 + h^4)(h^4 + H^4 + \|p^n - P_H^n\|_{M,H}^2) \\ &\leq C(H^{4-d} + H^{-d}\Delta t^2 + h^4)(h^4 + H^4 + \Delta t^2) \\ &\leq C(H^{8-d} + h^4 + \Delta t^2 + H^{-d}h^4\Delta t^2 + H^{-d}\Delta t^4). \end{aligned} \quad (5.75)$$

Combining equation (5.71) with equations (5.73)-(5.75), taking the $\delta\|\boldsymbol{\eta}^n\|^2$ term to the left side, multiplying by $2\Delta t$ and summing over $n, n = 1, \dots, N^*$ where N^* is the time step at which $\|\gamma^n\|$ achieves its maximum value gives,

$$\begin{aligned}
& \|\gamma^{N^*}\|^2 - \|\gamma^0\|^2 + \Delta t \sum_{n=1}^{N^*} K_* \|\boldsymbol{\eta}^n\|_{\text{TM}}^2 \\
& \leq \Delta t \sum_{n=1}^{N^*} (\|\epsilon^n\|^2 + \|d_t \alpha^n\|^2) + C \Delta t \sum_{n=1}^{N^*} \|\gamma^n\|^2 \\
& \quad + C(H^{4-d} + H^{-d}\Delta t^2 + h^4)\Delta t \sum_{n=1}^{N^*} (\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{\text{TM},H}^2 + \|\tilde{\mathbf{u}}^n - \tilde{\mathbf{U}}_h^n\|_{\text{TM}}^2 + H^4) \\
& \quad + C(h^4 + \|\gamma^{N^*}\|^2)\Delta t \sum_{n=1}^{N^*} H^{-d} \|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|^2 \\
& \quad + C(H^{8-d} + h^4 + \Delta t^2 + H^{-d}h^4\Delta t^2 + H^{-d}\Delta t^4).
\end{aligned}$$

Recalling the bound on ϵ^n , using Theorem 5.2 and recalling the initial conditions on \underline{P}_h^0 and P_h^0 gives,

$$\begin{aligned}
\|\gamma^{N^*}\|^2 + \Delta t \sum_{n=1}^{N^*} K_* \|\boldsymbol{\eta}^n\|_{\text{TM}}^2 & \leq C(H^{8-d} + h^4 + \Delta t^2 + H^{-d}h^4\Delta t^2 + H^{-d}\Delta t^4) \\
& \quad + C\Delta t \sum_{n=1}^{N^*} \|\gamma^n\|^2 + \tilde{C}\|\gamma^{N^*}\|^2(H^{4-d} + H^{-d}\Delta t^2).
\end{aligned}$$

We can choose H and Δt such that $H^{4-d} + H^{-d}\Delta t^2 \leq \frac{1}{2\tilde{C}}$, and the last term can be moved to the left-hand side. Applying Gronwall's Lemma gives

$$\|\gamma^{N^*}\|^2 + \Delta t \sum_{n=1}^{N^*} K_* \|\boldsymbol{\eta}^n\|_{\text{TM}}^2 \leq C(H^{8-d} + h^4 + \Delta t^2).$$

Applying Theorem 5.2 and the triangle inequality gives the desired result. \square

5.1.3 Extensions to Multiple Levels

The above analysis carries through for multiple levels. In this case, the nonlinear problem would still be solved once on the coarsest grid. However, one could have multiple fine grids. On each of these finer and finer grids, the nonlinear term is expanded about the next coarser solution and the resulting linear system is solved. Adding more grids corresponds to adding more Newton-like iterations with each iteration taking place on the next finer grid.

5.2 A Two-Level Method for Richards' Equation

In this section, we consider applying these two-level ideas to Richards' equation. We discuss only the expanded mixed method and not the superconvergent finite difference case.

In order to get a sufficient rate of convergence on the coarse grid to transfer to the fine grid, we make the assumption of strictly partially saturated flow. Under this assumption, the scheme in Section 4.2 can be applied on the coarse grid giving optimal convergence. Thus, we turn to the fine grid.

Linearizations of the time derivative term lead to schemes which are nonconservative and give incorrect mass balance. Thus, we consider leaving the time term nonlinear and just linearizing the hydraulic conductivity, K .

Our discrete time fine grid scheme is to find $(P^n, \tilde{\mathbf{U}}^n, \mathbf{U}^n) \in (W_h, \mathbf{V}_h, \mathbf{V}_h)$ for each time step $n = 1, \dots, N$ satisfying,

$$(d_t \theta(P^n), w) + (\nabla \cdot \mathbf{U}^n, w) = (f^n, w), \quad (5.76)$$

$$(\tilde{\mathbf{U}}^n, \mathbf{v}) = (P^n, \nabla \cdot \mathbf{v}), \quad (5.77)$$

$$(\mathbf{U}^n, \mathbf{v}) = (K(P_H^n) \tilde{\mathbf{U}}^n, \mathbf{v}) + (K'(P_H^n) \tilde{\mathbf{U}}_H^n (P^n - P_H^n), \mathbf{v}). \quad (5.78)$$

The following theorem gives the convergence behavior of this scheme.

Theorem 5.5 For $(P^n, \tilde{\mathbf{U}}^n, \mathbf{U}^n) \in (W_h, \mathbf{V}_h, \mathbf{V}_h)$ defined as in equations (5.76)-(5.78), we have,

$$\begin{aligned} \|p^N - P^N\| + \left(\sum_{n=1}^N \|K(P_H^n)^{1/2} (\tilde{\mathbf{U}}^n - \tilde{\mathbf{u}}^n)\|^2 \Delta t \right)^{1/2} \\ \leq C(h^{k+1} + \Delta t + H^{2k+2-d/2}). \end{aligned} \quad (5.79)$$

Proof Subtracting the numerical scheme from the variational formulation, equations (3.13)-(3.15), gives the error equations,

$$(d_t(\theta(p^n) - \theta(P^n)), w) + (\nabla \cdot (\Pi \mathbf{u}^n - \mathbf{U}^n), w) = (\epsilon^n, w), \quad (5.80)$$

$$(\hat{\mathbf{u}}^n - \tilde{\mathbf{U}}^n, \mathbf{v}) = (\hat{p}^n - P^n, \nabla \cdot \mathbf{v}), \quad (5.81)$$

and

$$\begin{aligned} (\Pi \mathbf{u}^n - \mathbf{U}^n, \mathbf{v}) &= (\Pi \mathbf{u}^n - \mathbf{u}^n, \mathbf{v}) - (K(P_H^n) \tilde{\mathbf{U}}^n, \mathbf{v}) \\ &\quad - (K'(P_H^n) \tilde{\mathbf{U}}_H^n (P^n - P_H^n), \mathbf{v}) + (K(p^n) \tilde{\mathbf{u}}^n, \mathbf{v}). \end{aligned} \quad (5.82)$$

Let $\gamma^n = \hat{p}^n - P^n$, $\boldsymbol{\eta}^n = \hat{\mathbf{u}}^n - \tilde{\mathbf{U}}^n$, and $\boldsymbol{\xi}^n = \Pi \mathbf{u}^n - \mathbf{U}^n$. Take $w = \gamma^n$ in (5.80), $\mathbf{v} = \boldsymbol{\xi}^n$ in (5.81) and $\mathbf{v} = \boldsymbol{\eta}^n$ in (5.82) and combine to get,

$$\begin{aligned} & (d_t(\theta(p^n) - \theta(P^n)), \gamma^n) - (K(P_H^n) \tilde{\mathbf{U}}^n, \boldsymbol{\eta}^n) \\ &= (\epsilon^n, \gamma^n) - (\Pi \mathbf{u}^n - \mathbf{u}^n, \boldsymbol{\eta}^n) + (K'(P_H^n) \tilde{\mathbf{U}}_H^n (P^n - P_H^n), \boldsymbol{\eta}^n) \\ & \quad - (K(p^n) \tilde{\mathbf{u}}^n, \boldsymbol{\eta}^n). \end{aligned} \quad (5.83)$$

Expanding $K(p^n)$ gives,

$$K(p^n) = K(P_H^n) + K'(P_H^n)(p^n - P_H^n) + \frac{K''(\alpha^n)}{2}(p^n - P_H^n)^2.$$

Rewriting (5.83) and using this expansion gives,

$$\begin{aligned} & (d_t(\theta(p^n) - \theta(P^n)), p^n - P^n) + (K(P_H^n) \boldsymbol{\eta}^n, \boldsymbol{\eta}^n) \\ &= (\epsilon^n, \gamma^n) + (\Pi \mathbf{u}^n - \mathbf{u}^n, \boldsymbol{\eta}^n) - (K(p^n)(\hat{\mathbf{u}}^n - \tilde{\mathbf{u}}^n), \boldsymbol{\eta}^n) \\ & \quad + ((\gamma^n + (p^n - \hat{p}^n))K'(P_H^n) \tilde{\mathbf{U}}_H^n, \boldsymbol{\eta}^n) + (K(P_H^n)(p^n - P_H^n)(\hat{\mathbf{u}}^n - \tilde{\mathbf{U}}_H^n), \boldsymbol{\eta}^n) \\ & \quad + \frac{1}{2}((p^n - P_H^n)^2 K''(\alpha^n) \hat{\mathbf{u}}^n, \boldsymbol{\eta}^n) + (d_t(\theta(p^n) - \theta(P^n)), p^n - \hat{p}^n). \end{aligned} \quad (5.84)$$

By Lemma 4.3, we have,

$$\begin{aligned} & (d_t(\theta(p^n) - \theta(P^n)), p^n - P^n) \\ & \geq \int_{\Omega} d_t \int_{P^n}^{p^n} (\theta(\varphi) - \theta(P^n)) d\varphi dx \\ & \quad - C\{(p^n - P^n)^2 + (p^{n-1} - P^{n-1})^2 + (\Delta t^n)^2\}. \end{aligned} \quad (5.85)$$

We bound the time discretization term as in section 4.2,

$$\|\epsilon^n\| \leq \left\| \frac{\partial^2 \theta}{\partial t^2} \right\|_{L^2(t^{n-1}, t^n); L^2} (\Delta t^n)^{1/2}. \quad (5.86)$$

Combining equations (5.84)-(5.86) and applying the Cauchy-Schwarz and the arithmetic-geometric mean inequalities, we have,

$$\begin{aligned} & \int_{\Omega} d_t \left(\int_{P^n}^{p^n} (\theta(\varphi) - \theta(P^n)) d\varphi \right) dx + (K(P_H^n) \boldsymbol{\eta}^n, \boldsymbol{\eta}^n) \\ & \leq C\{\|p^n - P^n\|^2 + \|p^{n-1} - P^{n-1}\|^2\} + C\{(\Delta t)^2 + \|\partial_{tt} \theta\|_{L^2(t^{n-1}, t^n); L^2}^2 \Delta t^n + \|\gamma^n\|^2\} \\ & \quad + (d_t(\theta(p^n) - \theta(P^n)), p^n - \hat{p}^n) + C\|\Pi \mathbf{u}^n - \mathbf{u}^n\|^2 + \epsilon \|\boldsymbol{\eta}^n\|^2 + C\|\hat{\mathbf{u}}^n - \tilde{\mathbf{u}}^n\|^2 \\ & \quad + C\|\boldsymbol{\eta}^n\| \|\gamma^n\| (\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{L^\infty} + \|\tilde{\mathbf{u}}^n\|_{L^\infty}) + C\|(p^n - \hat{p}^n) \tilde{\mathbf{U}}_H^n\|^2 \\ & \quad + C\|(p^n - P_H^n)(\hat{\mathbf{u}} - \tilde{\mathbf{U}}_H^n)\|^2 + C\|(p^n - P_H^n)^2 \hat{\mathbf{u}}^n\|^2. \end{aligned} \quad (5.87)$$

Let \hat{N} be the time index where $\max_n \|p^n - P^n\|$ occurs. Multiply (5.87) by Δt^n and sum on $n = 1, \dots, M$, where $M \leq N$ and $M \geq \hat{N}$.

The first left-hand side term becomes,

$$\begin{aligned} \sum_{n=1}^M \Delta t^n \int_{\Omega} d_t \left(\int_{P^n}^{p^n} (\theta(\wp) - \theta(P^n)) d\wp \right) dx \\ = - \int_{\Omega} \left(\int_{P^0}^{p^0} (\theta(\wp) - \theta(P^0)) d\wp \right) dx + \int_{\Omega} \left(\int_{P^M}^{p^M} (\theta(\wp) - \theta(P^M)) d\wp \right) dx \\ \geq -C \|p^0 - P^0\|^2 + C \|p^M - P^M\|^2, \end{aligned} \quad (5.88)$$

where we have used Lemma 4.3.

By summation by parts,

$$\begin{aligned} \sum_{n=1}^M (d_t(\theta(p^n) - \theta(P^n)), p^n - \hat{p}^n) \Delta t^n \\ = - \sum_{n=1}^{M-1} (\theta(p^n) - \theta(P^n), d_t(p^{n+1} - P^{n+1})) \Delta t^n \\ - (\theta(p^0) - \theta(P^0), p^1 - \hat{p}^1) + (\theta(p^M) - \theta(P^M), p^M - \hat{p}^M), \end{aligned} \quad (5.89)$$

where we have used assumption (3.32) of a quasi-uniform time discretization. The first right-hand side term of equation (5.87) is bounded by,

$$\begin{aligned} \sum_{n=1}^M \{ \|p^n - P^n\|^2 + \|p^{n-1} - P^{n-1}\|^2 \} \Delta t^n \\ \leq 2 \sum_{n=1}^M \|p^n - P^n\|^2 \Delta t^n + \|p^0 - P^0\|^2 \Delta t. \end{aligned} \quad (5.90)$$

So, noting that $\|\gamma^n\| \leq \|p^n - \hat{p}^n\| + \|p^n - P^n\|$, we have,

$$\begin{aligned} \|p^M - P^M\|^2 + \sum_{n=1}^M (K(P_H^n) \boldsymbol{\eta}^n, \boldsymbol{\eta}^n) \Delta t^n \\ \leq C \sum_{n=1}^M \|p^n - P^n\|^2 \Delta t^n + \|p^0 - P^0\|^2 + C(\Delta t)^2 \\ + C \sum_{n=1}^M \|\hat{p}^n - p^n\|^2 \Delta t^n + \epsilon \sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \\ + C \sum_{n=1}^M \|\theta(p^n) - \theta(P^n)\| \|d_t(p^{n+1} - \hat{p}^{n+1})\| \Delta t^n \end{aligned}$$

$$+ C \sum_{n=1}^m \{ \|\Pi \mathbf{u}^n - \mathbf{u}^n\|^2 + \|\hat{\mathbf{u}}^n - \tilde{\mathbf{u}}^n\|^2 \} \Delta t^n + \sum_{i=1}^4 T_i. \quad (5.91)$$

We bound T_1 by.

$$\begin{aligned} T_1 &= C \sum_{n=1}^M \|\boldsymbol{\eta}^n\| \|\gamma^n\| (\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{L^\infty} + \|\tilde{\mathbf{u}}^n\|_{L^\infty}) \Delta t^n \\ &\leq C \sum_{n=1}^M \|\boldsymbol{\eta}^n\| \|p^n - P^n\| (\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{L^\infty} + \|\tilde{\mathbf{u}}^n\|_{L^\infty}) \Delta t^n \\ &\quad + C \sum_{n=1}^M \|\boldsymbol{\eta}^n\| \|\hat{p}^n - p^n\| (\|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{L^\infty} + \|\tilde{\mathbf{u}}^n\|_{L^\infty}) \Delta t^n \\ &\leq C \|p^{\hat{N}} - P^{\hat{N}}\| \left(\sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \right)^{1/2} \left(\sum_{n=1}^M \|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{L^\infty}^2 \Delta t^n \right)^{1/2} + \epsilon \sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \\ &\quad + C \|\hat{p} - p\|_{L^\infty(L^2)} \left(\sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \right)^{1/2} \left(\sum_{n=1}^M \|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|_{L^\infty}^2 \Delta t^n \right)^{1/2} \\ &\quad + C \sum_{n=1}^M \|p^n - P^n\|^2 \Delta t^n + C \sum_{n=1}^M \|p^n - \hat{p}^n\|^2 \Delta t^n \\ &\leq C \|p^{\hat{N}} - P^{\hat{N}}\| \left(\sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \right)^{1/2} (H^{-d/2}(H^{k+1} + \Delta t)) \\ &\quad + C \|\hat{p} - p\|_{L^\infty(L^2)} \left(\sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \right)^{1/2} (H^{-d/2}(H^{k+1} + \Delta t)) \\ &\quad + C \sum_{n=1}^M \|p^n - P^n\|^2 \Delta t^n + C \sum_{n=1}^M \|p^n - \hat{p}^n\|^2 \Delta t^n, \end{aligned} \quad (5.92)$$

where we have used the inverse assumption Theorem 3.1 and Theorem 4.2.

Now,

$$\begin{aligned} &\|p^{\hat{N}} - P^{\hat{N}}\| \left(\sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \right)^{1/2} (H^{-d/2}(H^{k+1} + \Delta t)) \\ &\leq \frac{1}{2} \|p^{\hat{N}} - P^{\hat{N}}\|^2 + \frac{1}{2} \left(\sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \right) (H^{-d/2}(H^{k+1} + \Delta t))^2. \end{aligned} \quad (5.93)$$

We choose H and Δt such that,

$$H^{-d/2}(H^{k+1} + \Delta t) \leq K_*^{1/2}. \quad (5.94)$$

Note that in two dimensions, this requires $k \geq 0$ and in three dimensions, we must have $k \geq 1$. Then, the second right-hand side term in (5.93) is bounded by,

$$\frac{K_*}{2} \sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n. \quad (5.95)$$

Similarly,

$$\begin{aligned} & \|\hat{p} - p\|_{L^\infty(L^2)} \left(\sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n \right)^{1/2} (H^{-d/2}(H^{k+1} + \Delta t)) \\ & \leq \frac{1}{2} \|\hat{p} - p\|_{L^\infty(L^2)}^2 + \frac{K_*}{2} \sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n. \end{aligned} \quad (5.96)$$

Thus, T_1 is bounded by,

$$T_1 \leq C \sum_{n=1}^M \|p^n - P^n\|^2 \Delta t^n + C \sum_{n=1}^M \|p^n - \hat{p}^n\|^2 \Delta t^n + \frac{K_*}{2} \sum_{n=1}^M \|\boldsymbol{\eta}^n\|^2 \Delta t^n. \quad (5.97)$$

The term T_2 is bounded by,

$$\begin{aligned} T_2 &= C \sum_{n=1}^M \|(p^n - \hat{p}^n) \tilde{\mathbf{U}}_H^n\|^2 \Delta t^n \\ &\leq C \sum_{n=1}^M (\|p^n - \hat{p}^n\|^2 \|\tilde{\mathbf{u}}^n\|_\infty^2 + \|p^n - \hat{p}^n\|_\infty^2 \|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|^2) \Delta t^n \\ &\leq C h^{2(k+1)} + C \sum_{n=1}^M h^{2(k+1)} \|\tilde{\mathbf{U}}_H^n - \tilde{\mathbf{u}}^n\|^2 \Delta t^n, \end{aligned} \quad (5.98)$$

and T_3 by,

$$\begin{aligned} T_3 &= C \sum_{n=1}^M \|(p^n - P_H^n)(\hat{\mathbf{u}}^n - \tilde{\mathbf{U}}_H^n)\|^2 \Delta t^n \\ &\leq C \sum_{n=1}^M \|p^n - P_H^n\|_\infty^2 \|\hat{\mathbf{u}}^n - \tilde{\mathbf{U}}_H^n\|^2 \Delta t^n \\ &\leq C \sum_{n=1}^M (H^{-d/2}(H^{k+1} + \Delta t))^2 \|\hat{\mathbf{u}}^n - \tilde{\mathbf{U}}_H^n\|^2 \Delta t^n, \end{aligned} \quad (5.99)$$

where we have again used the inverse assumption.

Lastly,

$$T_4 = C \sum_{n=1}^M \|(p^n - P_H^n) \hat{\mathbf{u}}^n\|^2 \Delta t^n$$

$$\begin{aligned}
&\leq C \sum_{n=1}^M \|p^n - P_H^n\|_\infty^2 \|p^n - P_H^n\|^2 \|\hat{\mathbf{u}}^n\|_\infty^2 \Delta t^n \\
&\leq C \sum_{n=1}^M (H^{-d/2}(H^{k+1} + \Delta t))^2 (H^{k+1} + \Delta t)^2 \Delta t^n.
\end{aligned} \tag{5.100}$$

Thus, again using Theorem 4.2 we have for these last terms,

$$\begin{aligned}
\sum_{i=2}^4 |T_i| &\leq C(h^{2(k+1)} + (H^{-d/2}(H^{k+1} + \Delta t))^2 (H^{k+1} + \Delta t)^2 \\
&\quad + C \sum_{n=1}^M \|p^n - P^n\|^2 \Delta t^n + C \sum_{n=1}^M \|p^n - \hat{p}^n\|^2 \Delta t^n.
\end{aligned} \tag{5.101}$$

Now, let $M = \hat{N}$. Applying approximation properties, Gronwall's Lemma 3.1 and taking $P^0 = \hat{p}^0$ in (5.91) gives,

$$\begin{aligned}
&\|p^{\hat{N}} - P^{\hat{N}}\|^2 + \sum_{n=1}^{\hat{N}} (K(P_H^n) \boldsymbol{\eta}^n, \boldsymbol{\eta}^n) \Delta t^n \\
&\leq C\{\Delta t^2 + H^{2(k+1)} + h^{2(k+1)} + H^{-d} H^{4(k+1)}\}.
\end{aligned} \tag{5.102}$$

So, $\|p^{\hat{N}} - P^{\hat{N}}\| \leq C\{\Delta t + h^{k+1} + H^{-d/2+2(k+1)}\}$.

Combining these equations and taking $M = N$ in (5.91) gives,

$$\begin{aligned}
&\|p^N - P^N\| + \left(\sum_{n=1}^N \|K(P_H^n)^{1/2} \boldsymbol{\eta}^n\|^2 \Delta t \right)^{1/2} \\
&\leq C(h^{k+1} + \Delta t + H^{2k+2-d/2}).
\end{aligned} \tag{5.103}$$

The triangle inequality finishes the result. \square

Chapter 6

Implementation and Numerical Results

In this chapter the implementation of a C++ three-dimensional Parallel Richards' Equation Solve code, PREQS, is described. We first discuss the equation formulation and discretization scheme used. Then, a brief discussion of the nonlinear and linear discrete system solution techniques is given. Lastly, some numerical test cases are presented along with results.

6.1 Implementation Issues

In order to achieve parallelism, the parallelepiped domain, Ω , is decomposed into a set of smaller parallelepiped subdomains, denoted by Ω_i , so that $\Omega = \cup_i \Omega_i$. The data corresponding to the cells in each of these subdomains is stored on a single processor. Let $\Gamma_i = \partial\Omega_i$ and $\Gamma_{I_i} = \Gamma_i \setminus \Gamma$. Thus, Γ_{I_i} is the part of the boundary of subdomain i that is contained in Ω .

As shown in Section 5.1, the expanded mixed finite element method with certain quadrature rules simplifies to a cell-centered finite difference scheme with a 19 point stencil. This stencil is shown in Figure 6.1. In order to implement this finite difference scheme in parallel, each processor communicates with up to 18 neighbors. To reduce this requirement, we add extra unknowns along the interfaces between subdomains. Adding these unknowns allows for the normal fluxes at the interface points to be discontinuous, which is a nonphysical condition. Thus, extra equations which enforce continuity of normal fluxes at the interfaces are also introduced. Adding these extra unknowns corresponds to adding a single hydraulic head at the interface points. This value will be “owned” by one processor and communicated to the “non-owner” after updates.

Formulating a mixed method with these extra unknowns along interfaces between subdomains was first done by Glowinski and Wheeler [35] in the context of linear elliptic equations. They used these unknowns to formulate domain decomposition schemes for the mixed method.

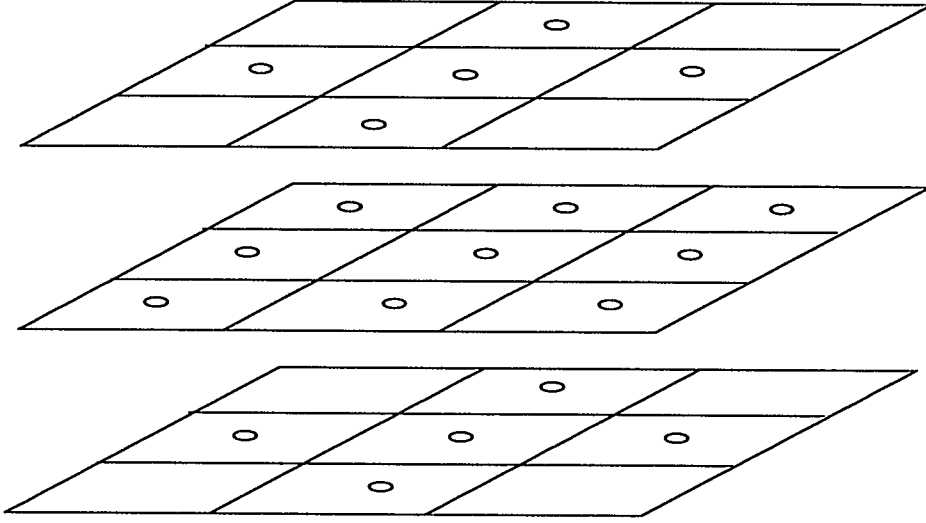


Figure 6.1 The 19-point discretization stencil.

General conditions of the form,

$$\sigma \mathbf{u} \cdot \mathbf{n} + vp = g,$$

are considered for the external boundaries. The functions σ and v are assumed to be functions of position only. The data g is a function of both time and position.

To define the numerical scheme, we define a new discrete vector space, $\hat{\mathbf{V}}_h$. Let $\mathbf{V}_i = \mathbf{V}_h|_{\Omega_i}$. Then take, $\hat{\mathbf{V}}_h = \bigoplus \mathbf{V}_i$. The numerical scheme is defined as finding $(P^n, \tilde{\mathbf{U}}^n, \mathbf{U}^n, L^n) \in (W_h, \hat{\mathbf{V}}_h, \hat{\mathbf{V}}_h, \Lambda_h)$ at each time step $n = 1, \dots, N$ satisfying,

$$(d_t \theta(P^n), w) + (\nabla \cdot \mathbf{U}^n, w) = (f^n, w), \quad (6.1)$$

$$(\tilde{\mathbf{U}}^n, \mathbf{v})_{\Omega_i, TM} = (P^n, \nabla \cdot)_{\Omega_i} - (L^n, \mathbf{v} \cdot \mathbf{n})_{\Gamma_i}, \quad (6.2)$$

$$(\mathbf{U}^n, \mathbf{v})_{\Omega_i, TM} = (K(P^n) \tilde{\mathbf{U}}^n, \mathbf{v})_{\Omega_i, T}, \quad (6.3)$$

$$(\sigma \mathbf{U}^n \cdot \mathbf{n}, \beta)_{\Gamma_i \cap \Gamma} = (g + v L^n, \beta)_{M, \Gamma_i \cap \Gamma}, \quad (6.4)$$

$$\sum_i (\mathbf{U}^n \cdot \mathbf{n}, \beta)_{\Gamma_i} = 0. \quad (6.5)$$

The extra unknowns provide a boundary condition for internal interfaces. Thus, the subdomains are coupled only through shared boundaries, and each subdomain will only need to communicate with neighbors sharing interfaces. Hence, subdomains communicate with up to 6 neighbors and not 18.

The main disadvantage of this approach is that as the number of subdomains increases, so does the number of extra unknowns. Thus, the algorithm changes as more processors are added, and parallel speedup, in the traditional sense, will be non-optimal.

In calculating flow velocities, the technique of one point upstream weighting [46] is employed. For this technique, the relative permeability and its derivative at interfaces are approximated by the value at the cell one point upstream. The upstream point is defined in the case of a full permeability tensor and cell-centered finite differences by [21],

$$k_{rw}(P_{i+1/2}) = \begin{cases} k_{rw}(P_{i+1}), & \text{if } k(x)\tilde{\mathbf{U}}_{i+1/2} \leq 0, \\ k_{rw}(P_i), & \text{otherwise.} \end{cases} \quad (6.6)$$

Forsyth and Kropinski [32] and Sammon [61] have shown that upstream weighting is necessary in order to accurately track the fluid front. However, by Taylor's Theorem the truncation error associated with upstream weighting is $O(h)$. Thus, we expect this approximation to be only $O(h)$.

We use an inexact Newton method [23] with backtracking line search globalization [24, 27] to solve the discrete nonlinear problem. In this method the Jacobian system is solved inexactly, most often with an iterative method. The GMRES Krylov subspace method [60] preconditioned with a Jacobi preconditioner is used in the code. The linear system tolerances are chosen using an algorithm of Eisenstat and Walker [28] which prevents oversolving of the system. Far from the solution, the nonlinear function F and its local linear approximation may disagree significantly. In this case, forcing the linear solution to be very accurate may lead to a step which provides little or no progress toward a solution. Furthermore, solving the linear system to a high level of accuracy can be very costly. Eisenstat and Walker give a variety of choices for tolerances [28]. One choice is,

$$\eta^m = \frac{|\|F(P^m)\| - \|F(P^{m-1}) + J(P^{m-1})\Delta P^m\||}{\|F(P^{m-1})\|},$$

where P^m is the current Newton iterate. This choice reflects the agreement between the function and its linear model at the previous step. Eisenstat and Walker have shown that for this choice of η^m , once the iterates are close enough to the solution, the inexact Newton method shows two-step quadratic convergence. This method has been effectively implemented in a two-phase flow code where the compute time decreased significantly with this choice for the linear system tolerances [21].

The PREQS code uses the kScript scripting language of Keenan [42] in order to provide a flexible user interface. The kScript command generation program cmdGen [41] was used to define kScript commands and variables. As a result, code input can be set in any units and in any order. Furthermore, Keenan has developed extensive array and vector C++ classes that were used throughout the PREQS code.

6.2 Numerical Results

In this section, results are given for the PREQS code applied to various test problems. A three-dimensional nonlinear heat equation with a known solution is considered first. Then, one-dimensional and three-dimensional partially saturated flow problems are discussed.

6.2.1 A Known Solution Test Case

In order to verify the asymptotic rate of convergence, a test problem with a known closed form analytical solution is considered. For this case, θ and k_r are both taken to be p , giving the equation,

$$\frac{\partial p}{\partial t} - \nabla \cdot (pk \nabla p) = f, \quad (6.7)$$

where f is chosen so that $p = xyz + 1$. The problem domain is taken to be the unit cube and,

$$k = \begin{bmatrix} 1.0 & 0.1 & 0.1 \\ 0.1 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{bmatrix}.$$

Dirichlet boundary conditions are taken on all sides of the domain and the initial condition is 1 everywhere. The nonlinear iteration tolerance was set to 10^{-9} , and all problems were solved on a processor mesh of $2 \times 2 \times 1$. Time steps of 0.001 were taken, and the discrete L^2 error measured at the final time, 0.1. The time step was taken to ensure that $\Delta t \leq h^2$ for all h considered.

Table 6.1 gives the discrete L^2 error for various mesh sizes. Let E_i be the error after solving on a grid of size h_i . Then, $E_i = Ch_i^r$ is the discrete L^2 error at the last time step. Taking logs results in,

$$(\log h_i)r + \log C = \log E_i. \quad (6.8)$$

Writing this for each data pair (E_i, h_i) gives an overdetermined set of equations for the convergence rate r and the log of the constant C . Using Matlab to solve this system, we arrive at the solution $r = 1.9$, $\log C = -6.3411$. This line, along with the five data points, is plotted in Figure 6.2.

Table 6.1 Convergence results for a three dimensional analytic test problem.

Grid	L^2 error
$2 \times 2 \times 2$	4.2947×10^{-4}
$4 \times 4 \times 4$	1.3104×10^{-4}
$8 \times 8 \times 8$	3.4952×10^{-5}
$16 \times 16 \times 16$	8.8112×10^{-6}
$32 \times 32 \times 32$	2.1521×10^{-6}

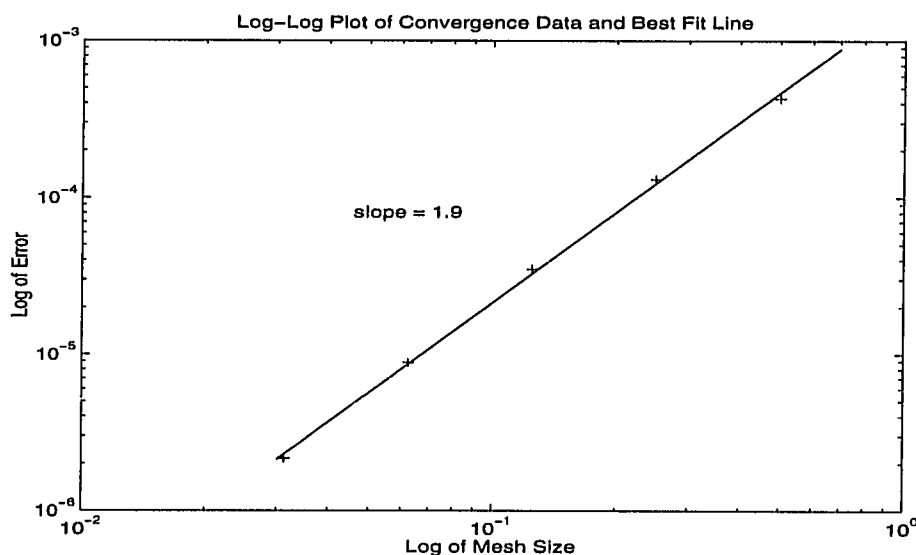


Figure 6.2 Linear plot of convergence data.

The analysis in Chapter 4 indicates that the expanded mixed method with the lowest order space should give at least $O(h)$ spatial convergence. However, the above results indicate that it may be possible to prove better than $O(h)$. Furthermore, the code uses a finite difference scheme based on superconvergent points similar to the

coarse grid scheme proven to be $O(h^2)$ for the nonlinear heat equation in Chapter 5. The method used in the PREQS code employs one-point upstream weighting which is $O(h)$ instead of the $O(h^2)$ bilinear interpolation. However, for problems where no steep fronts are present, it is not surprising to see better than $O(h)$ convergence even though upstream weighting is used.

6.2.2 A One-Dimensional Flow Problem

The next test problem we consider was reported by Celia, Bouloutas and Zarba [18] and is a one-dimensional flow problem. This problem was implemented in order to verify conservation of mass and to examine the effect of lengthening time steps.

Physical properties for this test case are given in Table 6.2. The water con-

Table 6.2 Physical data for the one-dimensional flow problem.

Domain	1 cm \times 1 cm \times 60 cm
α	0.0335
n	2
m	0.5
θ_s	0.368
θ_r	0.102
k_{xx}, k_{yy}, k_{zz}	1.0568×10^{-4} cm ²
k_{xy}, k_{xz}, k_{yz}	0
Density	1gm/cm ³
Viscosity	1.124cP
Porosity	0.368

tent and relative permeability are given by the van Genuchten curves, (2.2) and (2.7). Initial and boundary conditions for pressure head were as follows, $h(z, 0) = -1000$ cm, $h(0, t) = -1000$ cm and $h(60\text{cm}, t) = -75$ cm. No flow boundary conditions were taken on the four remaining boundaries. The depth direction was divided into cells of width 2.5cm, and a single processor was used for all results with this test case.

Let W denote the time change of water content in the domain over the time of simulation, and let F denote the water mass entering the domain over the time of simulation. Then,

$$W \equiv \sum_{n=1}^N \sum_i \frac{\theta_i^n - \theta_i^{n-1}}{\Delta t^n} \Delta t^n \Delta z = \sum_i (\theta_i^N - \theta_i^0) \Delta z,$$

$$F \equiv \sum_{n=1}^N \left(\sum_i (u_{i+1/2}^n - u_{i-1/2}^n) \cdot \mathbf{n}_i \right) \Delta t^n = \sum_{n=1}^N (u_0^n \cdot \mathbf{n}_0 - u_B^n \cdot \mathbf{n}_B) \Delta t^n,$$

where $u_B^n \cdot \mathbf{n}$ is the flux on the $z = 60\text{cm}$ boundary. The mass balance ratio is given by,

$$MB \equiv \frac{W}{F}. \quad (6.9)$$

If this ratio is unity, the numerical method exactly conserves mass.

Celia, et.al. show the mass balance ratio for a head-based form of Richards' equation solved with both a Galerkin finite element method and a finite difference method. Both schemes show large degradations of the mass balance as the time step increases. For steps of 1 minute, the finite difference scheme gives a ratio close to 1, but as Δt goes to 60 minutes, the ratio drops to 0.6. The finite element scheme gives even worse results. However, with the mixed form of Richards' equation, mass should be conserved and the ratio should be close to unity.

This test problem was solved with the PREQS code for one simulation day with various time steps ranging from 1 minute to 60 minutes. In all cases, the above mass balance ratio was always unity. Thus, no water was artificially created or destroyed by the numerical method.

Figure 6.3 shows the approximate solutions for four different time steps. These solutions are almost identical indicating that the mixed formulation of Richards' equation used in this work prevents degradation in results due to time step increases, unlike h -based forms which degrade quickly with step increases. This degradation can be dramatic and is documented in Celia, et.al. [18].

Figure 6.4 shows approximate solutions after 0.5 simulation days for a fixed time step of 15 minutes and varying spatial steps. The solution is converging to a sharper and sharper front indicating that as the grid is refined, the solution improves.

6.2.3 A Three-Dimensional Irregular Geometry Flow Problem

The last problem we consider is a three-dimensional problem over an irregular geometry domain.

For this case, the domain can be mapped to a rectangular domain of $100\text{cm} \times 100\text{cm} \times 20\text{cm}$ with a C^2 map, F . The theory of Arbogast, Wheeler and Yotov [7] discusses the transformation of the original problem to one over the rectangular computational domain. Specifically, if the mapping is C^2 , then the problem can

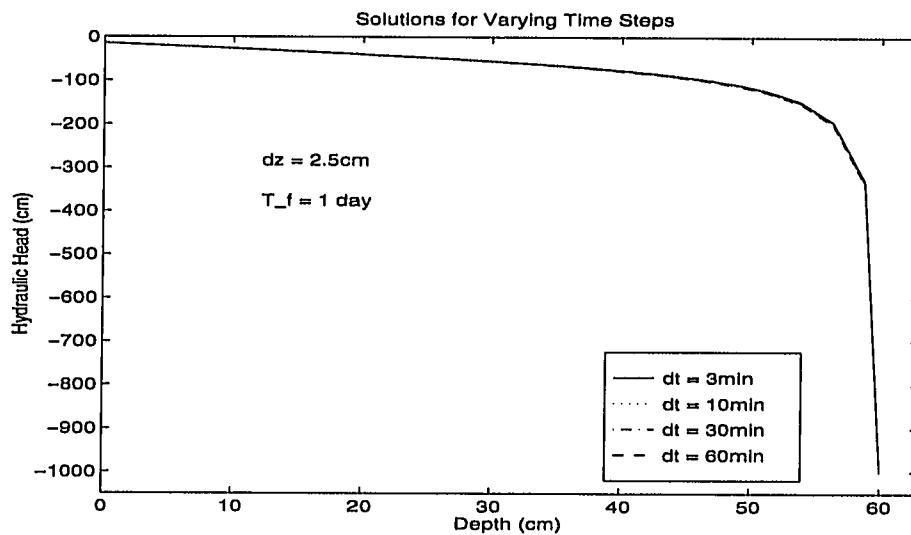


Figure 6.3 Solutions for the one-dimensional test problem with various time step sizes.

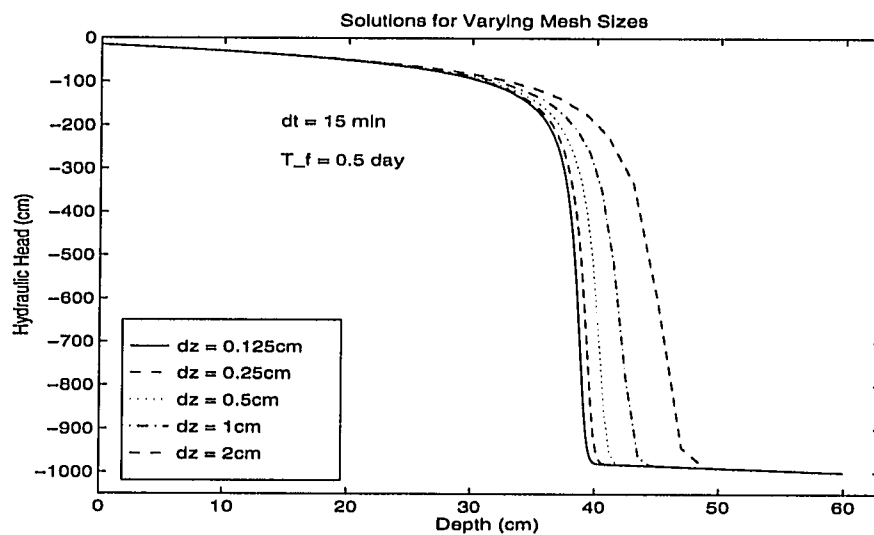


Figure 6.4 Solutions for the one-dimensional test problem with various mesh sizes.

be transformed into an equivalent problem over a rectangular domain which has a convergent solution. The permeability tensor, \mathcal{K} , is transformed by,

$$K = JDF^{-1}\mathcal{K}(DF^{-1})^T, \quad (6.10)$$

where DF is the Jacobian of the map, F , and J is the determinant of DF . The resulting permeability tensor will be full even in the case that the original is diagonal. Furthermore, the time derivative term is multiplied by J . Applying these transformations allows computation over a regular grid.

Physical properties for the original case are given in Table 6.3. The water content

Table 6.3 Physical data for the three-dimensional flow problem.

Domain	100 cm \times 100 cm \times 20 cm
α	0.0334
n	2
m	0.5
θ_s	0.361
θ_r	0.102
$k_{xx}, k_{yy}, k_{zz}, z \leq 10\text{cm}$	$9.33 \times 10^{-12} \text{ m}^2$
$k_{xy}, k_{xz}, k_{yz}, z \leq 10\text{cm}$	$9.33 \times 10^{-13} \text{ m}^2$
$k_{xx}, k_{yy}, k_{zz}, z > 10\text{cm}$	$9.33 \times 10^{-10} \text{ m}^2$
$k_{xy}, k_{xz}, k_{yz}, z > 10\text{cm}$	$9.33 \times 10^{-11} \text{ m}^2$
Density	1gm/cm ³
Viscosity	1.124cP
Porosity	0.368

and relative permeability are again given by the van Genuchten curves, (2.2) and (2.7). The computational grid was $20 \times 20 \times 10$ divided uniformly over a $2 \times 2 \times 1$ processor mesh. As seen in Table 6.3, the domain has two horizontal layers. The top layer has a much higher permeability than the lower. No flow boundary conditions were taken on all boundaries except the top and bottom. On the top face, the $x \in (0, 20\text{cm}), y \in (0, 20\text{cm})$ section had a hydraulic head boundary condition of -50 cm and no flow conditions everywhere else. On the bottom face, the $x \in (80\text{cm}, 100\text{cm}), y \in (0, 20\text{cm})$ section had a hydraulic head condition of -1000 cm and no flow everywhere else. These conditions effectively placed a source at the top left section of the domain and a sink at the bottom right section.

Figure 6.5 shows results for the regular domain in the case that the entire domain has the higher permeabilities given in Table 6.3 after 45 simulation days. The infiltrating water has advanced radially outward from the injection part of the upper boundary and has been pulled downward by gravity. The water has reached the sink boundary condition and has started flowing out of the domain. Figure 6.6 shows the regular domain for the two permeability layer case given in Table 6.3 after 50 simulation days. Here, we see that the water does not easily flow into the low permeability region. The water must accumulate enough weight in order to push into this region. Furthermore, the low hydraulic head condition is not felt by the water since it has not yet gotten to that part of the boundary.

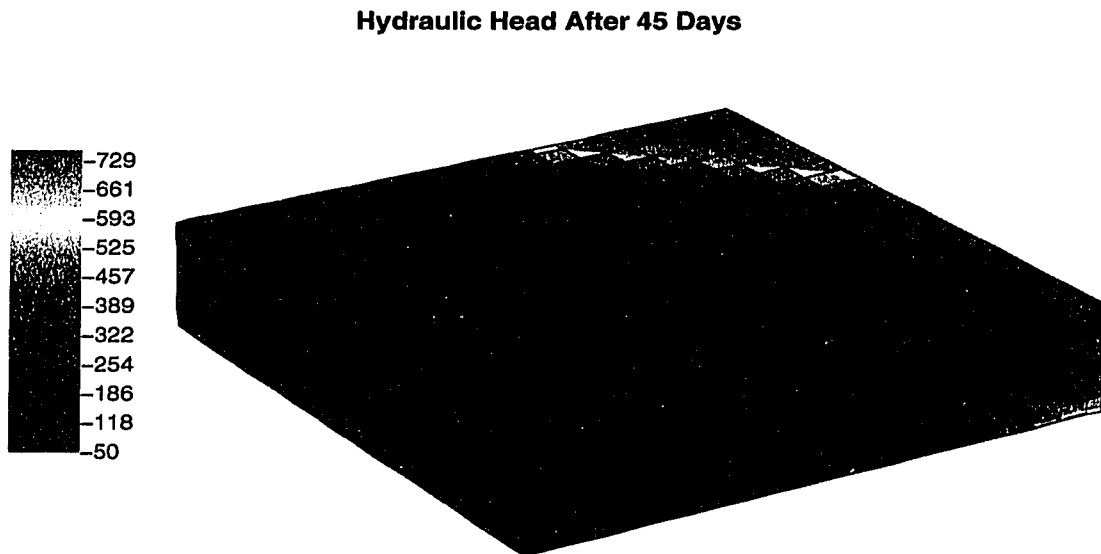


Figure 6.5 Three-dimensional single permeability layer test case after 45 simulation days.

Lastly, Figures 6.7-6.10 show the hydraulic head after 5, 20, 50 and 75 simulation days for the irregular geometry test case. As can be seen from the figures, water starts flowing at the top left of the domain. As time passes, it begins drifting toward the right while gravity pulls it downward. However, when the water reaches the lower permeability region, it must accumulate enough pressure to go further. Instead of going straight down, it pools along the interface between the two regions. As enough pressure builds underneath the injection area, the water is pushed downward and

Hydraulic Head After 50 Simulation Days

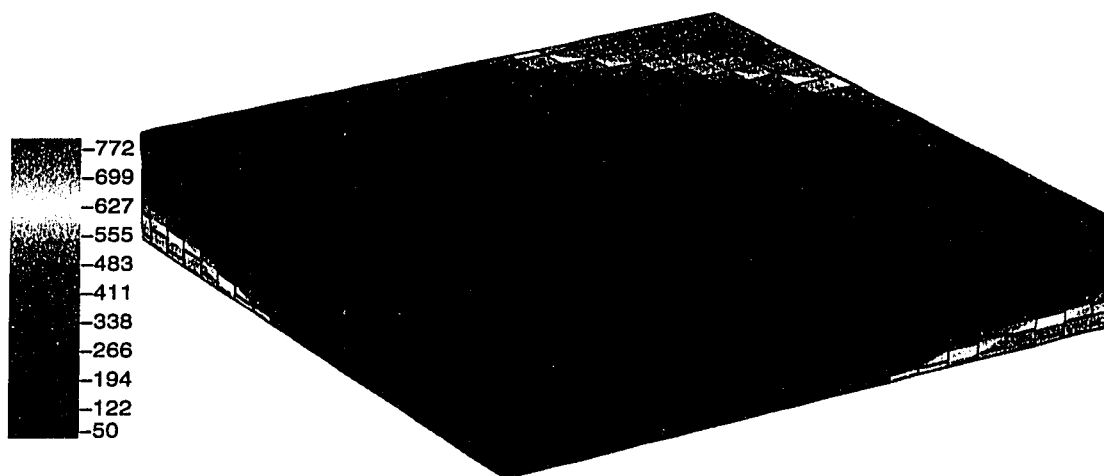


Figure 6.6 Three-dimensional two permeability layer test case after 50 simulation days.

begins to accumulate at the bottom of the domain. Around 20 days we can see the preference of the water to flow toward the rightmost part of the domain. This is due to the shape of the domain as it is twisted at the floor. Thus, the downward direction is toward the lower front of the domain. By 75 days, most of the upper region has filled with water and the water is starting to flow into the lower region. The effect of the two layers is clearly seen. Furthermore, it is clear that domain shapes have a dramatic effect on the flow and should not be modeled by simple rectangles.

In conclusion, we have shown that the PREQS code gives $O(h^2)$ accuracy on a model problem with a known solution. Furthermore, the code maintains perfect mass balance and is robust for large time steps. Lastly, the code predicts expected solutions to three-dimensional groundwater problems with full tensor coefficients.

Hydraulic Head After 5 Simulation Days

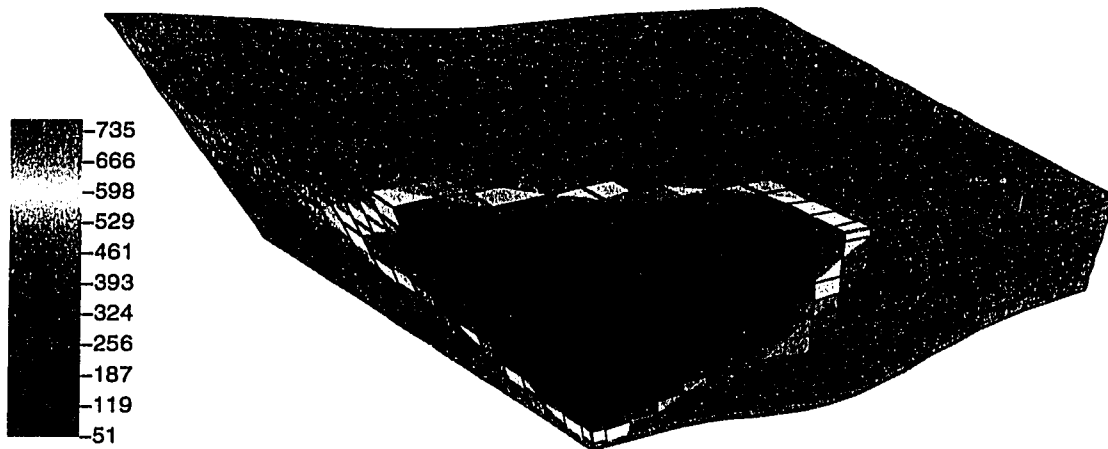


Figure 6.7 Three-dimensional irregular geometry test case after 5 simulation days.

Hydraulic Head After 20 Simulation Days

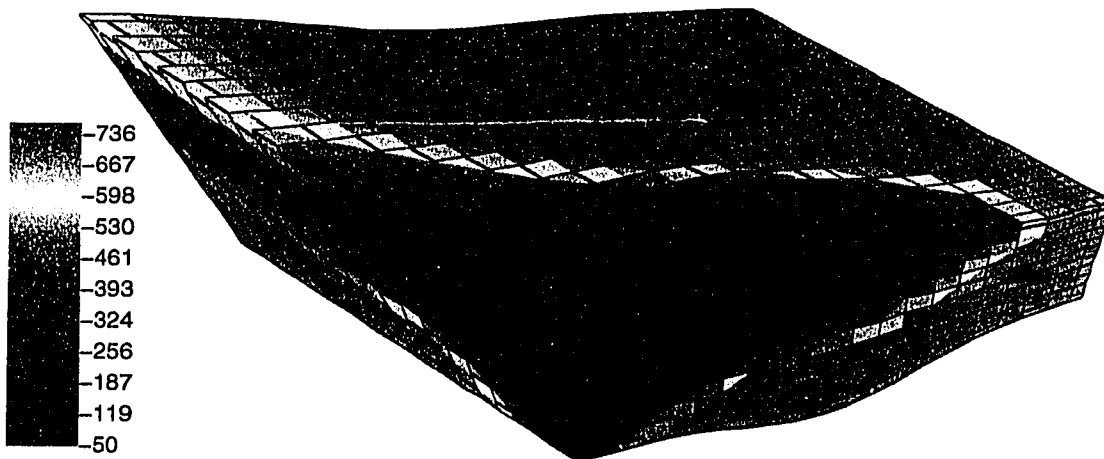


Figure 6.8 Three-dimensional irregular geometry test case after 20 simulation days.

Hydraulic Head After 50 Simulation Days

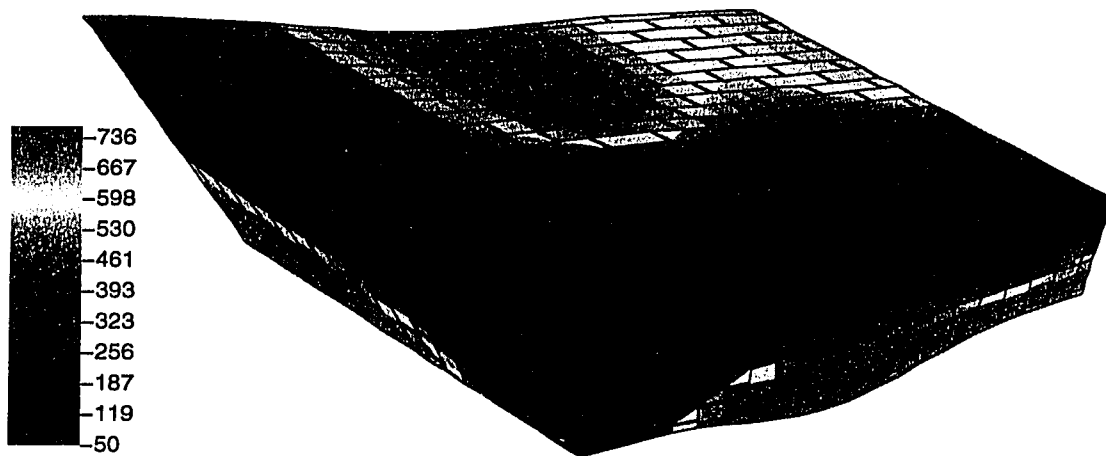


Figure 6.9 Three-dimensional irregular geometry test case after 50 simulation days.

Hydraulic Head After 75 Simulation Days

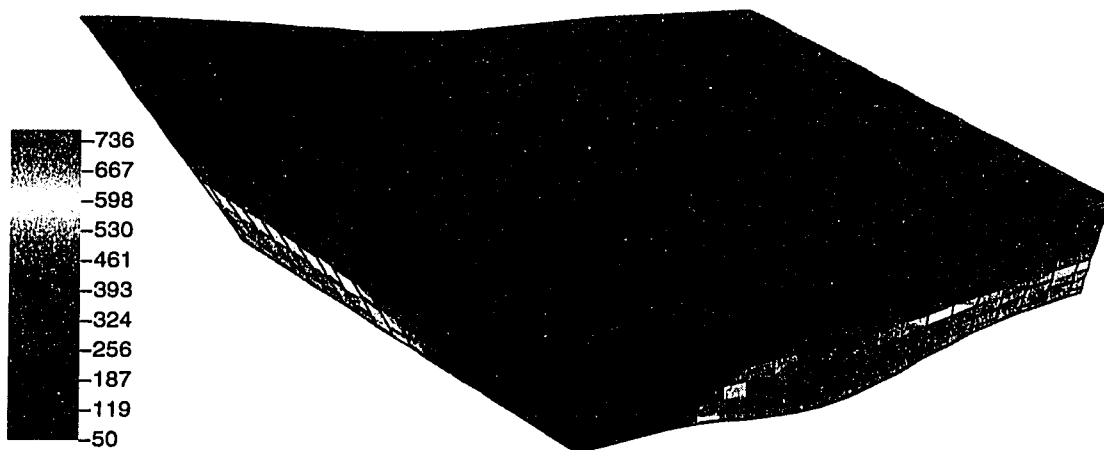


Figure 6.10 Three-dimensional irregular geometry test case after 75 simulation days.

Chapter 7

Conclusions

7.1 Summary

In this thesis we have analyzed and solved numerical methods for simulating variably saturated subsurface flow. A brief summary of existing literature in this area was presented along with the physical situation that is modeled.

We developed and analyzed a number of expanded mixed finite element methods applied to Richards' equation. Optimal convergence was shown for a discrete time scheme applied to the strictly variably saturated case. Optimal convergence was also shown for a nonlinear form in the case of partially to fully saturated flow. The nonlinear form is bounded below by the error in water content and above by the error in hydraulic head. Convergence rates in terms of a Hölder continuity rate were shown for the error in water content for this case. Lastly, convergence rates in terms of approximation error were shown for the unsaturated to fully saturated flow case with a continuous time scheme. This scheme was defined using the Kirchhoff transform in order to put the equation into a more easily analyzable form.

A method for handling nonlinearities was examined where the nonlinear problem is solved on a coarse grid and the problem on the fine grid linearized about the coarse grid solution. This technique is difficult to apply to Richards' equation due to the need to maintain mass balance. Thus, a scheme where the time derivative term on the fine grid was left nonlinear was analyzed.

Numerical results were given for a three-dimensional parallel Richards' equation solve code, PREQS. The code exhibited superconvergence on a model problem as well as robustness for large time steps. Mass balance is maintained exactly in the code. Results from a full tensor three-dimensional irregular geometry test case were shown and expected behavior was produced.

7.2 Future Work

Despite the above results, there is still plenty of work to be done in the area of simulating partially saturated subsurface flow.

The PREQS code uses only a Jacobi preconditioner and can be quite slow for long simulations where the linear systems get difficult to solve. Thus, future work will include finding robust preconditioners for the linear nonsymmetric Jacobian systems. Furthermore, the dynamic linear system tolerances described in Chapter 6 can significantly slow convergence of the Newton method. Further work will try to identify ways of making the forcing term selection more robust.

Richards' equation has been criticized as a simplistic model for expressing the two-phase flow of water and air [36]. However, when applicable, it is much faster to solve Richards' equation than a full two-phase model such as described in [21]. An interesting area for future work would be to closely examine under which numerical and physical conditions the solution of Richards' equation fails to match that of the full two-phase flow problem.

Bibliography

- [1] L. M. ABRIOLA AND J. R. LANG, *Self-adaptive hierarchic finite element solution of the one-dimensional unsaturated flow equation*, Int. J. Num. Meth. Fluids, 10 (1990), pp. 227–246.
- [2] M. B. ALLEN AND C. MURPHY, *A finite element collocation method for variably saturated flows in porous media*, Numerical Methods for Partial Differential Equations, 3 (1985), pp. 229–239.
- [3] H. W. ALT AND S. LUCKHAUS, *Quasilinear elliptic-parabolic differential equations*, Mathematische Zeitschrift, 183 (1983), pp. 311–341.
- [4] T. ARBOGAST, *An error analysis for Galerkin approximations to an equation of mixed elliptic-parabolic type*, Dept. Comp. Appl. Math. TR90-33, Rice University, Houston, TX 77251, Oct. 1990.
- [5] T. ARBOGAST, M. OBEYESKERE, AND M. F. WHEELER, *Numerical methods for the simulation of flow in root-soil systems*, SIAM J. Numer. Anal., 30 (1993), pp. 1677–1702.
- [6] T. ARBOGAST, M. F. WHEELER, AND NAI-YING ZHANG, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, Dept. Comp. Appl. Math. TR94-17, Rice University, Houston, TX 77251, Apr. 1994. To appear SIAM J. Num. Anal., 1996, vol. 33.
- [7] T. ARBOGAST, M. F. WHEELER, AND I. YOTOV, *Logically rectangular mixed methods for groundwater flow and transport on general geometry*, Dept. Comp. Appl. Math. TR94-03, Rice University, Houston, TX 77251, Jan. 1994.
- [8] —, *Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences*, Dept. Comp. Appl. Math. TR95-06, Rice University, Houston, TX 77251, Mar. 1995. To appear SIAM J. Numer. Anal., 1997, vol. 34.

- [9] D. N. ARNOLD AND F. BREZZI, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, Math. Model. and Numer. Anal., 19 (1985), pp. 7–32.
- [10] O. AXELSSON, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.
- [11] J. BEAR, *Dynamics of Fluids in Porous Media*, Elsevier, New York, 1972.
- [12] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
- [13] F. BREZZI, *On the existence, uniqueness and approximation of saddle point problems arising for lagrangian multipliers*, R.A.I.R.O. Anal. Numer., 8 (1974), pp. 129–151.
- [14] F. BREZZI, J. DOUGLAS, JR, R. DURÁN, AND M. FORTIN, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.
- [15] F. BREZZI, J. DOUGLAS, JR, M. FORTIN, AND L. D. MARINI, *Efficient rectangular mixed finite elements in two and three space variables*, Math. Modelling and Num. Ana., 21 (1987), pp. 581–604.
- [16] F. BREZZI, J. DOUGLAS, JR, AND L. D. MARINI, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.
- [17] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.
- [18] 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 (1990), pp. 1483–1496.
- [19] Z. CHEN, *Expanded mixed finite element methods for linear second order elliptic problems I*, IMA Preprint Series 1219, IMA, University of Minnesota, 1994.

- [20] P. J. DAVIS, *Interpolation and Approximation*, Dover Publications, New York, 1975.
- [21] C. N. DAWSON, H. KLIE, C. A. SAN SOUCIE, AND M. F. WHEELER, *A parallel, implicit, cell-centered method for two-phase flow*, Texas Inst. for Comp. and Applied Math., University of Texas, Austin, TX, 1996. Manuscript in preparation.
- [22] C. N. DAWSON AND M. F. WHEELER, *Two-grid methods for mixed finite element approximations of nonlinear parabolic equations*, Contemporary Mathematics, 180 (1994), pp. 191–203.
- [23] R. S. DEMBO, S. C. EISENSTAT, AND T. STEIHAUG, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
- [24] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1983.
- [25] R. G. DURÁN, *Superconvergence for rectangular mixed finite elements*, Numer. Math., 58 (1990), pp. 287–298.
- [26] L. J. DURLOVSKY, *Numerical calculation of equivalent grid block permeability tensors for heterogeneous porous media*, Water Resources Research, 27 (1991), pp. 699–708.
- [27] S. C. EISENSTAT AND H. F. WALKER, *Globally convergent inexact Newton methods*, SIAM J. Optimization, 4 (1994), pp. 393–422.
- [28] —, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comp., 17 (1996), pp. 16–32.
- [29] R. E. EWING, R. D. LAZAROV, AND J. WANG, *Superconvergence of the velocity along the gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.
- [30] G. FAIRWEATHER, *Finite Element Galerkin Methods for Differential Equations*, Marcel Dekker, Inc., New York, 1978.
- [31] C. W. FETTER, *Contaminant Hydrogeology*, Macmillan, New York, 1992.

- [32] P. A. FORSYTH AND M. C. KROPINSKI, *Monotonicity considerations for saturated-unsaturated flow*, Dept. of Comp. Sci. CS-94-17, University of Waterloo, 1994.
- [33] P. A. FORSYTH, Y. S. WU, AND K. PRUESS, *Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media*, *Advances in Water Resources*, 18 (1995), pp. 25–38.
- [34] R. A. FREEZE AND J. A. CHERRY, *Groundwater*, Prentice Hall, Inc., New Jersey, 1979.
- [35] R. GLOWINSKI AND M. F. WHEELER, *Domain decomposition and mixed finite element methods for elliptic problems*, in *Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, R. Glowinski et al., eds., SIAM, Jan. 1987, pp. 144–172.
- [36] W. G. GRAY AND S. M. HASSANIZADEH, *Paradoxes and realities in unsaturated flow theory*, *Water Resources Research*, 27 (1991), pp. 1847–1854.
- [37] R. HAVERKAMP AND M. VAUCLIN, *A comparative study of three forms of the Richard equation used for predicting one-dimensional infiltration in unsaturated soil*, *Soil Sci. Soc. of Am. J.*, 45 (1981), pp. 13–20.
- [38] M. T. VAN GENUCHTEN, *A closed form equation for predicting the hydraulic conductivity of unsaturated soils*, *Soil Sci. Soc. Am. J.*, 44 (1980), pp. 892–898.
- [39] R. G. HILLS, I. PORRO, D. B. HUDSON, AND P. J. WIERENGA, *Modeling one-dimensional infiltration into very dry soils 1. Model development and evaluation*, *Water Resour. Res.*, 25 (1989), pp. 1259–1269.
- [40] P. S. HUYAKORN, S. D. THOMAS, AND B. M. THOMPSON, *Techniques for making finite elements competitive in modeling flow in a variably saturated porous media*, *Water Resour. Res.*, 20 (1984), pp. 1099–1115.
- [41] P. T. KEENAN, *cmdGen 2.5 user manual*, Texas Inst. for Comp. and Applied Math. 96-09, University of Texas, Austin, TX, Feb. 1996.
- [42] ———, *kScript 2.5 user manual*, Texas Inst. for Comp. and Applied Math. 96-08, University of Texas, Austin, TX, Feb. 1996.

- [43] M. R. KIRKLAND, R. G. HILLS, AND P. J. WIERENGA, *Algorithms for solving Richards' equation for variably saturated soils*, Water Resour. Res., 28 (1992), pp. 2049–2058.
- [44] J. KOEBBE, *A computationally efficient modification of mixed finite element methods for flow problems with full transmissivity tensors*, Numer. Meth. for PDE's, 9 (1993), pp. 339–355.
- [45] O.-A. LADYZHENSKAYA, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969. English translation of Russian, 2nd Edition.
- [46] C. MATTAX AND R. DALTON, *Reservoir Simulation*, vol. 13, SPE-Monograph Series, Richardson, TX, 1990.
- [47] P. C. D. MILLY, *A mass-conservative procedure for time-stepping in models of unsaturated flow*, Advances in Water Resources, 8 (1985), pp. 32–36.
- [48] F. MILNER, *Mixed finite element methods for quasilinear second-order elliptic problems*, Math. of Comp., 44 (1985), pp. 303–320.
- [49] M. NAKATA, A. WEISER, AND M. WHEELER, *Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains*, in The Mathematics of Finite Elements and Applications V, Academic Press, Inc., London, 1985.
- [50] J. NEDELEC, *Mixed finite elements in \mathbb{R}^3* , Numer. Math., 35 (1980), pp. 315–341.
- [51] R. H. NOCHETTO AND C. VERDI, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
- [52] J. M. ORTEGA AND W. C. REINBOLT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, San Diego, 1970.
- [53] K. RATHFELDER AND L. ABRIOLA, *Mass conservative numerical solutions of the head-based Richards equation*, Water Resources Research, 30 (1994), pp. 2579–2586.

- [54] P. A. RAVIART AND J. M. THOMAS, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods: Lecture Notes in Mathematics 606, I. Galligani and E. Magenes, eds., Berlin, 1977, Springer-Verlag, pp. 292–315.
- [55] L. A. RICHARDS, *Capillary conduction of liquids through porous mediums*, Physics, 1 (1931), pp. 318–333.
- [56] J. E. ROBERTS AND J.-M. THOMAS, *Mixed and hybrid finite element methods*, Rapports de Recherche 737, INRIA, Oct. 1987.
- [57] M. E. ROSE, *Numerical methods for flows through porous media. i*, Mathematics of Computation, 40 (1983), pp. 435–467.
- [58] P. J. ROSS AND K. L. BRISTOW, *Simulating water movement in layered and gradational soils using the Kirchhoff transform*, Soil Sci. Soc. of Am. J., 54 (1990), pp. 1519–1524.
- [59] T. F. RUSSELL AND M. F. WHEELER, *Finite element and finite difference methods for continuous flows in porous media*, in Mathematics of reservoir simulation, R. E. Ewing, ed., SIAM, Philadelphia, 1983, ch. II, pp. 35–106.
- [60] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
- [61] P. H. SAMMON, *An analysis of upstream weighting*, Soc. Pet. Eng. J. Res. Eng., 3 (1988), pp. 1053–1056.
- [62] A. WEISER AND M. WHEELER, *On convergence of block-centered finite differences for elliptic problems*, SIAM J. Numer. Anal., 25 (1988), pp. 351–357.
- [63] J. XU, *Two-grid finite element discretizations for nonlinear elliptic equations*, Dept. of Mathematics AM105, Pennsylvania State University, University Park, Pennsylvania, July 1992. To appear in SIAM J Numer. Anal.
- [64] —, *A novel two-grid method for semilinear elliptic equations*, SIAM J. Sci. Comput., 15 (1994), pp. 231–237.