R88-08

THREE-DIMENSIONAL FLOW IN RANDOM POROUS MEDIA

Vol. 1

by
RACHID ABABOU
LYNN W. GELHAR
and
DENNIS McLAUGHLIN

RALPH M. PARSONS LABORATORY
HYDROLOGY AND WATER RESOURCE SYSTEMS

Report Number 318

Prepared under the support of the
Nuclear Regulatory Commission, Contract NRC-04-83-174
National Science Foundation, Grant ECE-8311786

March, 1988

MIT
DEPARTMENT
OF
CIVIL
ENGINEERING
SCHOOL
MASSACHUSETTS
Cambridge,

ERRATUM:

Vol. 1, p.2 (Abstract):

Second sentence of last paragraph of p.2 should read:

"Porous medium heterogeneity inferred from borehole conductivity data is represented by self-similar random fields. The flow equation is solved for small scale fluctuations" (etc. . . .).

Vol. 1, p.386 (Table 5.4):

On the row corresponding to values of the relaxation parameter ω, the first value is 0.25, not 2.5.

Vol. 1, p.389 (Eq. 5.110 and below):

The term $C_0^{-1/2}$ should be replaced by $C_0^{+1/2}$ in eq. (5.110). In addition, $C_0^{1/2}$ should be replaced by $C_0^{-1/2}$ two lines below eq. (5.110); the convergence rate r is proportional to $C_0^{-1/2}$, not $C_0^{+1/2}$.

Vol. 2, Chap. 7, Figures 7.13-14-15 (pp.662-665):

The contour labels on Figure 7.13, Figure 7.14, and bottom plot of Figure 7.15 are erroneous. These labels should be changed as follows:

(1) 70 → 125

(2) 60 → 100

(3) 50 → 90

(4) 40 → 80

(5) 30 → 70

The positive numbers indicate suction head in centimeters (not pressure head).

THREE-DIMENSIONAL FLOW IN RANDOM POROUS MEDIA

by

Rachid Ababou
Lynn W. Gelhar
Dennis McLaughlin

Large scale subsurface flow in heterogeneous porous formations is studied using a three-dimensional (3D) random field representation of local hydraulic properties, assuming for the most part that the medium is homogeneous/ergodic. Both analytical and numerical methods are used to characterize the physical behaviour of 3D flow fields, based on the statistical properties of solutions of stochastic partial differential equations.

The first order spectral theory is applied to the case of flow in stratified aquifers, revealing a nearly isotropic head correlation structure in vertical planes, in contrast with the anisotropy of conductivity and velocity. Thus, the infinite domain spectral solutions may only be applicable to aquifers much deeper than the vertical head correlation scale, say tens of horizontal conductivity scales. Also, non-perturbative spectral equations derived from mass conservation and statistical symmetries relate the flux and head spectra for isotropic media, independently of small parameter expansions. The new relations are satisfied by the current spectral solutions in any dimension. In the 2D case, we establish a statistical identity between head gradient and flux, or head and stream function (statistical conjugacy). Other results in the literature inspired a simple conjecture for the effective conductivity tensor with arbitrary 3D anisotropy of the random conductivity field, which fits all results known to be exact. Finally, a modified flux spectrum is developed using an approximate spectral solution of equations governing the flux instead of head. The new expressions for velocity variance and solute macrodispersivity appear to follow more realistic behaviours at high variability.

The conceptual approach of "spectral conditioning" is developed to describe finite size effects, particularly for evolving subsurface phenomena. Porous medium heterogeneity inferred from borehole conductivity data is solved for small scale fluctuations up to domain scale, conditioned on larger scale fluctuations (effective variability versus uncertainty). Closed-form results for 1D flow show that the uncertainty of finite domain statistics, such as effective conductivity and head variance, decreases with domain size. Preliminary results also indicate the scale dependence of the 3D

effective conductivity and macrodispersivity of a growing contaminant plume.

Direct numerical simulations are developed for single realizations of homogeneous/ergodic random medium properties, using the 3D turning band method for random field generation, and a special purpose numerical code for solving 3D finite difference saturated and unsaturated flow equations, with spatially variable and nonlinear coefficients. In the case of stochastic groundwater flow, a novel method of truncation error analysis shows that the discrete solution is a consistent approximation of the exact one, however with a lower order of accuracy if the random coefficients are noisy rather than smooth. The root-mean-square errors on head and velocity are respectively proportional to the powers 3/2 and 1/2 of the mesh-correlation scale ratio in the noisy case (exponential covariance). The convergence of the iterative "strongly implicit procedure" solver is studied, and the nonlinear stability of transient unsaturated flow systems is analyzed (Peclet number condition).

The single realization approach is applied to the following flow problems, using a Cray2 supercomputer for the largest simulations: steady groundwater flow in statistically isotropic or anisotropic media (up to 1 million nodes); transient strip source infiltration and steady rainfall infiltration in unsaturated soils with random conductivity-pressure curves (up to 30,000 nodes). The simulated flow fields are statistically analyzed by spatial averaging methods under weak assumptions of homogeneity. For groundwater flow, the results compare favourably with spectral solutions, especially for the head variance, effective conductivity, and velocity correlation tensor, up to large standard deviation of the natural logarithm of conductivity (2.3). The numerical velocity variances agree with the spectral theory at moderate variability (isotropic case), but increase faster with conductivity variance. The discrepancy is milder for the new flux-based spectral theory. Numerical head correlations tend to be smaller than theoretical ones due to finite size effects, particularly for shallow stratified aquifers.

For unsaturated flow, simulation results indicate sensitivity of flow behaviour depending on the variability and anisotropy of the random soil. In the case of transient strip source infiltration/drainage in a statistically anisotropic soil, there is a pronounced lateral spreading of the edges of the moisture plume. This behaviour is in qualitative agreement with available spectral solutions. The case of steady "rainfall" infiltration shows a quantitative agreement with the head variance and vertical unsaturated effective conductivity from the spectral theory. Some questions remain open, notably concerning the range of validity of the homogeneity and ergodicity hypotheses for highly nonlinear and evolving flow systems.

ACKNOWLEDGEMENTS

This research was supported in part by the Nuclear Regulatory Commission (Contract NRC-04-83-174) and the National Science Foundation (Grants ECE-8311786 and ECE-8544457). Computational support was provided by Parsons Laboratory MicroVax facility, the Minnesota Supercomputer Center and the NASA-Ames Research Center.

This research was part of a team effort and has benefited at times from the contributions of Andrew Tompson, Wendy Graham, and Don Polmann, to name a few. Conversations with many other students and faculty, both within the hydrology group of Parsons Laboratory and outside, also helped shape this research. We wish to thank in particular Tom Nicholson of the Nuclear Regulatory Commission for coordinating and enlivening our meetings.

This report is essentially a reproduction of the thesis submitted in January 1988 by the first author, Rachid Ababou, in partial fulfillment of the requirements for the degree of Doctor of Philosophy. Lynn W. Gelhar, Professor of Civil Engineering, served as thesis advisor. While he was on sabbatical leave, the work was guided by Dennis McLaughlin, Associate Professor of Civil Engineering. Additional members of the thesis committee were Michael A. Celia (Department of

Civil Engineering), Benoit B. Mandelbrot (IBM Thomas J. Watson Research Center), and Anthony T. Patera (Department of Mechanical Engineering); their comments and suggestions on this research are gratefully acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Subsurface Contamination and Field Heterogeneity

There has been significant progress during the past decade in our conceptual understanding of the physics of flow and mass transport in naturally heterogeneous or random porous media. Stochastic concepts were introduced to represent the natural variability of porous (non-fractured) subsurface formations and soils, leading to new mathematical formulations of the classical equations describing subsurface flow and transport phenomena, which in turn influenced current stochastic approaches to data collection.

This area of research has gradually grown into a field of its own, known as "stochastic subsurface hydrology". The basic ingredients of this approach are, on the mathematical side, the theory of random functions of multidimensional space (random fields) and of stochastic partial differential equations; the concept of "effective" transport coefficients (macro-scale conductivity and dispersitivity); and advanced linear estimation/optimization theory for the collection of noisy field data. Similar concepts have been used in the past, notably in the statistical theory of homogeneous turbulence (random velocity fields) and in the mathematical theory of "homogeneization",

where interest focuses on the existence and uniqueness of effective transport properties for a variety of physical problems.

The field of stochastic subsurface hydrology has become an active area of research for applications to toxic and radioactive waste contamination, with the increasing public and governmental awareness of the gravity and ubiquity of the toxic waste problem (North America and Western Europe). Because water is the main carrier of toxic species underground, e.g., in dissolved form, the study of subsurface water flow is an essential preliminary step towards a better understanding of toxic solute transport in complex environments.

The complexity of natural subsurface flow systems can be reduced somewhat by considering separately two distinct types of flow regimes: purely saturated flow with positive water pressures (aquifers), and purely unsaturated flow with negative water pressures below atmospheric pressure (unsaturated soils and vadose zone). In this work, we will be mainly concerned with the physics of water flow under these two distinct regimes, with particular emphasis on the effects of random-like heterogeneities of the porous medium.

[a] Subsurface contamination:

It may be useful to consider briefly how research on heterogeneous flow systems relates to actual field contamination problems. Low-level radioactive or toxic chemicals are usually disposed of above the ground or buried at shallow depths. Examples of hazardous waste sites of this type are landfills, surface impoundments (e.g. lined evaporation ponds), and uranium mill tailings buried at shallow depths. Another situation of interest concerns the case of potentially harmful chemicals applied over large surfaces at relatively small concentrations, such as may occur in irrigated areas (fertilizers dissolved in irrigation water). In these cases, contaminant transport will presumably take place in the unsaturated flow regime when a leak occurs. Except for controlled experiments, leakage is usually not detected until after the contaminant has reached a major, extensively monitored groundwater system, or until it has caused major damage (e.g. contamination of local communities via drinking water).

Large-scale controlled experiments of contaminant migration through the vadose zone are scant (see the extensive review of Gelhar et. al. 1984). One notable exception is the experimental study of an evaporation pond by Trauntwein et. al. (1983) and Kent et. al. (1982). These authors found that, after

20 years of leakage in the unsaturated zone, water reached a depth of 100 meters, and extended laterally over a distance of 2 kilometers. The surprisingly large lateral spread could be explained by the presence of horizontal clay lenses, leading perhaps to the formation of perched water zones (Gelhar et. al., 1984). The flow system in this case is complex and inherently three-dimensional in nature.

Another type of application concerns the case of high-level radioactive wastes. Currently, a number of options are being (re)considered for their disposal, particularly in the United States. One of these options is the deep burial of sealed cannisters into saturated formations of very low permeability, such as *unfractured* basalt, tuff, or granite. Another option which has been proposed, but not yet implemented to our knowledge, is the burial of radioactive wastes into very dry unsaturated porous formations located in arid (and poorly populated) regions. Winograd (1981), one of the proponents of this option, discusses the possibility of burying high-level wastes at depths of 15 m to 85 m in valley-fill deposits inside the man-made Sedan crater at Yucca Flat, Nevada. In that particular environment the annual rainfall is only 125 mm/year (5 inches/year) and the water table is 600 meters deep. Winograd estimated that the downward percolation rate (velocity) through

the unsaturated valley-fill could be roughly on the order of 2 mm/year, i.e., 200 meters per hundred thousand years. Incidentally, this gives also an idea of the large time scale of interest for high-level, slowly decaying radioactive wastes. In this case of very slow flow and extremely dry conditions, it is not at all clear what the effect of local heterogeneities could be on the overall pattern of dispersion of a contaminant at the scale of thousands to several hundred thousand of years.

Of more immediate concern is the potential hazard from existing waste facilities. For instance, a major leak was detected in 1973 beneath one of the tanks located at the storage facility of Hanford, Washington state. The total amount that leaked into the vadose zone was evaluated as one fifth of the 500,000 gallons (2000 m^3) of high-level radioactive liquid waste initially contained in the tank. The contaminant movement at this site is now being extensively monitored. In 1978, significant radionuclide concentrations were detected as far as 30 meters in depth, and 25 meters from the edge of the tank, laterally (Rouston, et. al., 1979).

Considering all the possible options and scenarios of subsurface contamination, it is always possible that, in case of a leak, the contaminant will eventually reach a major (regional) groundwater flow system, where transport takes place in the

saturated flow regime. There have been intense monitoring of contamination plumes in a number of groundwater systems worldwide. However, the groundwater velocities are generally not directly accessible to measurements. Thus, to *predict* the fate of contaminants in groundwater flow systems requires a priori a reliable modelization of groundwater velocities based on other types of information more accessible to measurement. The current practice is a subtle and complex combination of numerical modeling and model fitting, e.g., from measurement of hydraulic heads and concentrations on the site. For unsaturated flow and transport, however, there have been very few observations of the flow and contamination patterns at the large scale. The report by Gelhar et. al. (1984) contains a comprehensive review and interpretation of field data for both types of flow/transport problems, saturated and unsaturated. A major conclusion from their review is that models based on a simple extrapolation of small scale data often fail to predict phenomena occurring over large space-time scales.

[b] Field heterogeneity and implications:

One essential feature of subsurface formations, on which we have chosen to focus in this dissertation, is spatial variability. With the increasing body of literature devoted to the measurement and identification of various properties of

natural formations, it appears clear that heterogeneities occur at many scales, from "grain size" up to large geological structures. For most practical purposes, the micro-scale or "grain size" variations of a discontinuous nature can be filtered out. The formation is then viewed as a porous continuum, whose bulk properties reflect implicitly the granular nature of the medium (e.g. porosity, storage capacity, conductivity, and coefficient of mechanical dispersion of tagged fluid particles).

It should be recognized, however, that this simplified view may not always be realistic for soils with fissures or macropores, for fractured rocks, or for karstic formations. Such irreducibly discontinuous media will be ignored in this work. On the other hand, there is now ample evidence that the bulk properties of natural porous media, even without the presence of fractures or other large scale discontinuities, may vary quite erratically in space. It has been progressively recognized that such variability plays a major role in phenomena like solute dispersion at the large scale (macrodispersion).

Figure (1.1a) from Gelhar (1976) depicts the vertical variation of the log-saturated conductivity measured from small cores taken from a borehole. The conductivity fluctuates, apparently at random, over four orders of magnitude. It is

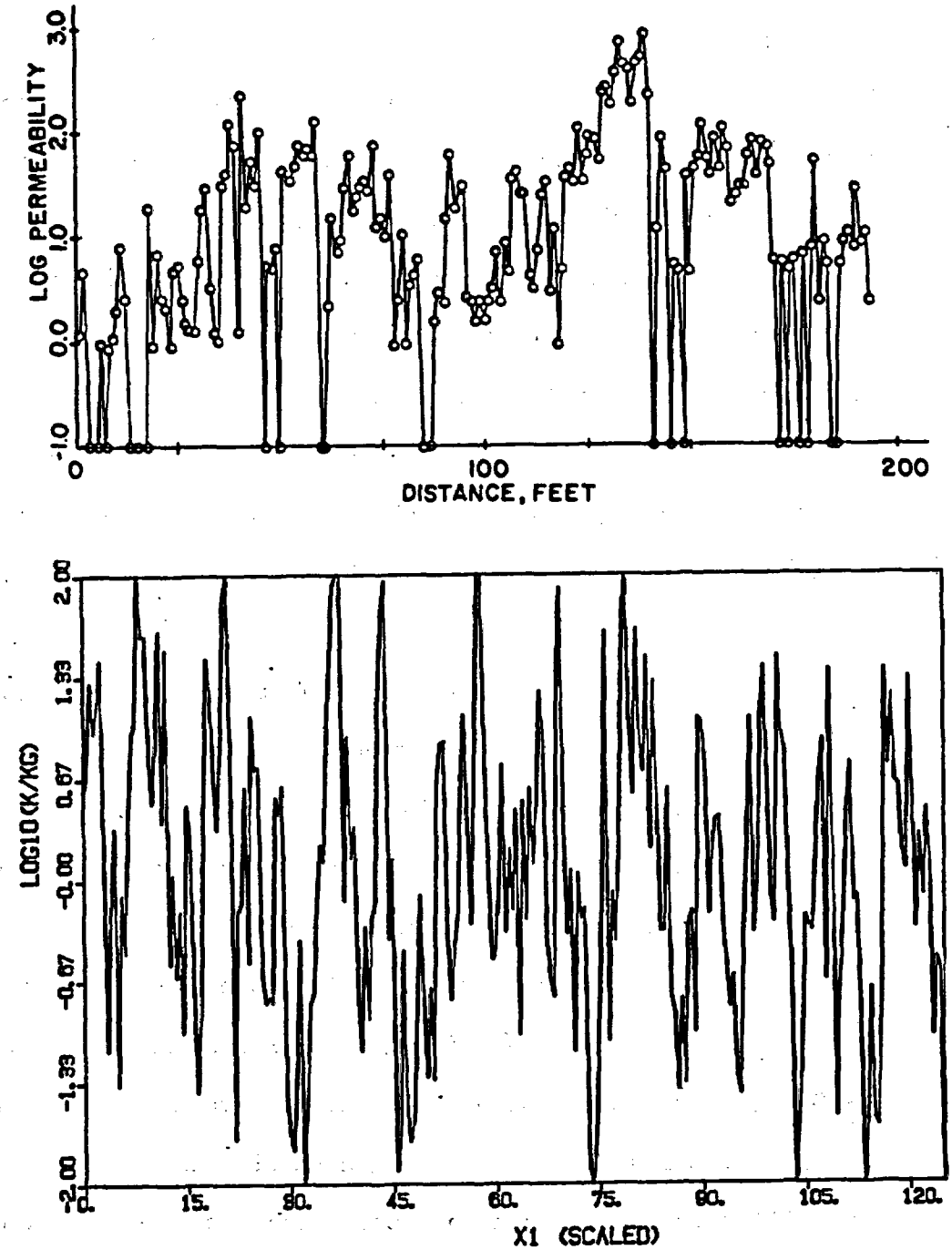


Figure 1.1: Examples of log-conductivity records $\ell n\, K(x)$ in one spatial dimension: (a) data from vertical borehole (Gelhar, 1976), (b) Synthetic realization of a one-dimensional Gauss-Markov process with exponential covariance function.

instructive to compare these data to the generated random function depicted in Figure (1.1b). The variance and correlation scale of the synthetic random function in (b) were adjusted to fit those of the observed record in (a). In many other cases as well, random functions appear to be adequate models of natural variability. Relevant data for aquifers can be found for example in papers by Freeze (1975), Delhomme (1979), and Hoeksema and Kitanidis (1985). Their work indicates that the log-saturated conductivity (ℓnK) is normally distributed, with standard deviations ranging from 0.5 to 3.5 (see review in Chapter 2).

The hydraulic properties of unsaturated soils also exhibit seemingly random variation in space. In this case, however, spatial variability is more difficult to characterize because unsaturated soil conductivity and moisture content are functions of soil water pressure. These functional relations are usually measured either on small soil samples in the laboratory, or on small field plots (1-10 m2, with a vertical resolution of 10-20 cm). Parameters such as porosity, saturated conductivity, and shape factors appear to be log-normally distributed whenever variability is high enough to distinguish a skewed distribution. Experimental evidence supporting this assertion can be found in Nielsen et. al. (1973), Warrick et. al. (1977), Sharma et. al. (1980), Gelhar et. al. (1982), and Russo (1983). A review of

available field data will be presented in Chapter 2.

In summary, there is ample evidence of seemingly random variations of hydraulic properties in space. Moreover, data like those shown in Figure (1.1) seem to exhibit a definite spatial structure (scale of fluctuation) that could be identified as the correlation length of a stationary random function. We recognize however that the correlation scale may not be uniquely identifiable in practice. In some experimental studies, the correlation scale was found negligible (or ignored altogether, as in Nielsen et. al., 1973). In others, it was found almost as large as the region under investigation, i.e. statistically meaningless. Finally, studies like those of Scisson and Wierenga (1981), Gajem et. al. (1981), and others, seem to indicate sensitivity of the observed variance and correlation scale with respect to the scale of measurement and size/spacing of measurement network. This, we feel, is an important issue that deserves a more thorough discussion. The topic will be touched upon at other places in this work.

1.2 Scope and Objectives:

[a] Motivation and background:

The present work is based, for the most part, on the

premise that a realistic description of naturally heterogeneous porous media can be achieved by representing the local hydraulic properties in the form of statistically homogeneous random fields in three-dimensional space. Previous research under this assumption have led to useful statistical characterizations of the global behavior and spatial structure of heterogeneous flow and solute transport. In particular, the spectral theory developed by Gelhar and co-workers (Gelhar, 1984 and 1987; Gelhar and Axness, 1983) has led to closed form relations describing in a compact form the statistical behavior of heterogeneous groundwater flow systems, including the cases of statistically isotropic and statistically anisotropic (stratified) aquifers. The global/statistical properties of interest for applications are the effective conductivity, the macrodispersion of a convected solute, and the degree of variability and spectral content of the random head and velocity fields. Parallel results have also been obtained by other stochastic approaches, which will be reviewed in Chapter 2.

The spectral theory provides particularly simple results due to the assumptions of infinite-domain and statistical homogeneity/ergodicity of the solutions of the stochastic flow and transport equations. These key assumptions, by reducing further the complexity of the problem, permitted to obtain *approximate* expansion solutions of the stochastic groundwater

flow equation (similar to a "heat equation" with random conductivities) and of the solute transport equation ("convection-diffusion equation" with a random velocity field) in terms of spectral densities in Fourier space. These spectral solutions give all the information there needs to know about the *covariance structure* of the variables of interest (hydraulic head and groundwater velocities). However, the validity of the approximations involved in the spectral theory has yet to be ascertained for a wide range of field conditions. The same remark holds for other analytical stochastic theories proposed in the literature. In general, the available stochastic solutions (spectral or other) rely on some kind of small parameter expansion, where the small parameter corresponds to the degree of variability of the underlying porous formation (e.g., standard deviation of the random log-conductivity field). In the case of the spectral theory, the accuracy of the small parameter expansion *and* the validity of the homogeneity/ergodicity hypothesis need to be checked for realistic situations.

Moreover, the application of stochastic concepts to the case of unsaturated flow in heterogeneous porous media has encountered serious difficulties, both technical and conceptual, due to the highly *nonlinear* nature of the hydraulic properties of unsaturated media (moisture capacity and hydraulic conductivity functions of water pressure). Although the results obtained by

the linearized spectral approach (Yeh et al. 1985, Mantoglou et. al. 1987) are qualitatively appealing, it is not at all clear yet what the range of applicability of their results could be. In view of our poor understanding of the complex interactions between nonlinear effects and spatial variability, there is a strong motivation for investigating the operational range of the linearized spectral theory of unsaturated flow in the vadose zone, notably concerning the predicted shape, anisotropy, and hysteresis of the unsaturated conductivity tensor as a function of mean water pressure. Even a more qualitative description of actual or *simulated* heterogeneous unsaturated flow patterns could be useful.

At the present date, the available data concerning the large scale pattern of heterogeneous unsaturated flow sytems are too scant (as noted previously) to allow for a verification of the linearized spectral results, except perhaps in a very qualitative and speculative fashion. On the other hand, there are now enough experimental observations to indicate that the findings of the spectral theory, among other stochastic approaches, may provide an adequate description of the variability of certain groundwater flow systems and its effect on contaminant macrodispersion (Gelhar, 1986). However, these indications still remain subjective, as too many undetermined parameters enter into play, notably the three-dimensional

correlation structure, fluctuation scales, and statistical anisotropy of the hypothetically random log-conductivity field.

Therefore, there arises the need for an independent verification of the analytical results of the spectral theory in more closely controlled situations, where the actual spatial variability of the underlying porous medium is known with reasonable accuracy.

[b] Present approach and objectives:

In this work, we aim at obtaining accurate solutions of the equations governing flow in hypothetically random porous media, based on the "postulate" of statistical homogeneity of the hydraulic properties of saturated or unsaturated media. In the case of *saturated groundwater flow* in particular, we will seek to refine and extend further the analytical results previously obtained by the first order spectral theory of Gelhar et al. There are several new approaches involved in the proposed refinements. In one instance, we explore higher order and non-perturbative solutions of the stochastic flow equation, while still retaining the infinite domain/ergodicity hypothesis. In another, we extend the spectral theory further to treat explicitly the influence of domain size on the statistical behavior of the flow system at some finite scale. However, the

latter approach will require the same approximation of "small variability" as in the infinite-domain approach.

On the other hand, we also develop and apply a *numerical* solution method in view obtaining, as accurately as possible, the solution of the three-dimensional saturated flow equation for finite discrete realizations of random medium properties. When the size of the domain is sufficiently large, the numerical solution of this equation should have the same statistical properties as predicted by the spectral theory if the latter was correct. Moreover, only one large single realization of the random medium should suffice to represent the ensemble statistics of the flow field, if the homogeneity/ergodicity hypothesis of the spectral theory holds true. Therefore, the numerical/single realization approach should provide an independent check of the accuracy of infinite-domain spectral solutions, and of its applicability in practical cases where the domain of investigation is necessarily finite.

Finally, the same numerical approach will be used to treat the case of three-dimensional unsaturated flow systems in random soils, i.e., with random coefficients intervening in the nonlinear constitutive properties (random unsaturated conductivity-pressure curve as a function of space). However, in this case, the approximations involved in the linearized

spectral theory are too severe, and the numerical requirements on the discrete grid size are too constraining, to allow for a precise statistical comparison between the spectral and single-realization approaches. Nevertheless, the large numerical experiments discussed in this work will appear useful for a preliminary screening of the complex behavior of heterogeneous unsaturated flow systems.

In summary, this dissertation focuses on the mathematical description of subsurface flow fields under the postulate that the natural variability of hydraulic properties can be characterized in the form of statistically homogeneous random fields in three-dimensional space. For the case of saturated flow, the variety of techniques that were used to solve the stochastic flow equation may reflect our attempts at minimizing the set of postulates and approximations required to achieve tractable results. On the other hand, our rather empirical approach of stochastic unsaturated flow reflects the difficulty of developing truly nonlinear yet tractable analytical models of heterogeneous flow systems in that case. Whenever possible, we have used the indications of the linearized spectral theory for interpretation. Finally, it may be relevant here to emphasize that, as in the case of turbulent flows, exact solutions to the saturated and unsaturated stochastic flow equations are not known. Thus, even approximate indications on

the robustness of existing theories can be extremely valuable.

1.3 Thesis Preview:

This dissertation was broken down into eight chapters, including this introductory part. The next chapter, *Chapter* 2, contains a brief literature review, a survey of available field data, and a general presentation of the single-realization approach, particularly in relation with the hypotheses of the spectral theory of stochastic flow. This chapter provides a background for subsequent developments. *Chapters* 3 *and* 4 are exclusively devoted to the obtention of tractable solutions of stochastic ground water flow by analytical means. *Chapter* 5 focuses on numerical issues related to the single-realization approach. This Chapter develops at length the various aspects of a saturated-unsaturated flow simulator, to be used as a tool in *Chapters* 6 *and* 7. The results of large saturated flow simulations in randomly heterogeneous media are presented in *Chapter* 6. The emphasis there is on the statistical analysis of the numerical flow fields for comparisons with the predictions of the spectral theory. *Chapter* 7 presents in a more qualitative fashion the results of large unsaturated flow simulations in random soils. The overall conclusions of this work are given in *Chapter* 8.

Because the amount of material presented in Chapters 3 - 7 is quite large, we have found it convenient to summarize the contents of each of these chapters in a compact form, as shown below:

Chapter 3:

The first order spectral theory of Gelhar and Axness (1983) is used to obtain a more complete picture of the statistical properties of the flow. Some new analytical results are derived by integrating the spectra to obtain ensemble moments, such as some closed form expressions for the variances of the flux vector components in the case of extreme anisotropy. A remarkable result concerns the near isotropy of the head field in statistically anisotropic media. The implications of these findings are discussed (shallow/deep stratified aquifers and finite size effects). The role of the low wavenumber fluctuations of the conductivity field is also briefly examined. The discussion focuses on the physical meaning and possible limitations of the spectral solutions.

Chapter 4:

In this chapter are developed several new analytical approaches related to the standard spectral theory of saturated flow in stochastic porous media. A new non-perturbative analysis

of the stochastic flow equation is developed independently of the "small variability" approximation of the spectral theory. In the restricted case of statistically isotropic conductivity, it is shown that the spectral theory of Gelhar and Axness (1983) satisfies all the fundamental conservation and symmetry properties of the flow in any number of dimensions. These properties lead to an exact statistical identity between the flux and head gradient vectors in the 2D isotropic case (statistical conjugacy). In the 3D isotropic case, the flux-head gradient relation contains a few undetermined functions. In the more general case of 3D anisotropic media, an extrapolation of previous results by Matheron (1967) and Gelhar and Axness (1983) leads to a simple closed form expression for the anisotropic effective conductivity tensor. A modification of the Gelhar-Axness spectral solutions to include higher order terms in the flux spectrum is developed, by using a stochastic system of equations governing the flux vector rather than the hydraulic head. Finally, we propose a generalization of the spectral approach to take into account the effects of finite domain size. The generalization is based on the new concept of "spectral conditioning", used to quantify explicitly the relative amounts of uncertainty and spatial variability with respect to the scale of the problem. The method combines ideas related to the concept of self-similarity (Mandelbrot, 1983), renormalization group methods (Wilson, 1975), and spectral solution method (Gelhar,

1984). Closed form results are obtained in particular for one-dimensional saturated flow. Some implications for three-dimensional flow and solute transport are also discussed in view of preliminary results.

Chapter 5:

This chapter is devoted to the development and analysis of a numerical method for solving large single realizations of saturated and unsaturated flow in three-dimensional random porous media. In the linear case, a new approach for evaluating finite difference truncation errors with stochastic coefficients is used to show that the classical centered finite difference scheme is consistent, at least in the mean-square sense. For the flux vector, the order of accuracy is equal to one for a smooth conductivity field, but drops to one half for a noisy field. The leading order terms of the head error and flux error are explicitly evaluated as functions of grid resolution for different kinds of random conductivity fields. We then focus on the practical implementation of an iterative matrix solver (SIP). Numerical experiments are presented for large random flow problems on the order of 0.1-1 million nodes (convergence rate analysis). An upper bound for the true solution error is given as a function of convergence rate and final residual error. The convergence rate appears proportional to the inverse square-root

of the condition number. Finally, we also develop a nonlinear system solver for unsaturated flow problems (nonlinear SIP). Preliminary numerical simulations of infiltration in some uniform and heterogeneous soil systems are used to demonstrate the problem solving capabilities of the unsaturated flow simulator. The conclusions regarding the numerical feasibility of the single realization approach are quite favorable in the saturated flow case. However, we also conclude that there could be some severe restrictions on the mesh size (Peclet number condition) and/or time step size in the case of transient unsaturated flow, particularly for dry heterogeneous soils.

Chapter 6:

This chapter is devoted to the interpretation and statistical analysis of large single realization simulations of three-dimensional saturated flow in random porous formations. The random conductivity fields are generated on the finite difference grid of the flow simulator by using the 3D turning band method. We begin with a preliminary analysis of "medium size" flow problems (130,000 nodes) in the case of statistically isotropic conductivities, with emphasis on the qualitative features of the head and flux fields, and some comparisons with the predictions of the spectral theory. We then move on to a more extensive statistical analysis of large flow simulations on

a 1 million node grid, again with statistically isotropic conductivities. The comparison with spectral results is quite favorable for a wide range of log-conductivity (ℓnK) standard deviations, up to $\sigma = 2.3$. The discrepancies observed for the flux variances are analyzed, and a further modification of the spectral solution is proposed to account for high order effects on flux variability. The agreement is quite good for other flow characteristics (head variability, flux correlation structure, effective conductivity) but the numerical head correlation ranges appear shorter due to finite size effects. We also investigate finite size effects for the case of statistically anisotropic media. The covariance structure of the head field is analyzed for two flow simulations mimicking the case of shallow and deep stratified aquifers with moderate anisotropy (grid size 220,000 nodes). The chapter ends with a summary of findings and conclusions on the range of validity of spectral solutions of stochastic groundwater flow.

Chapter 7:

This chapter presents a qualitative analysis of large single realization simulations of three-dimensional infiltration in random unsaturated soils. The random field coefficients of the exponential conductivity-pressure curve were generated by the turning band method. A preliminary analysis of strip source

infiltration for modest size realizations of statistically isotropic soils (25,000 nodes) revealed the effect of the variability of the slope of the conductivity-pressure curve and of its correlation with the saturated conductivity. Maximum variability of the moisture patterns was observed when both these parameters were random, and uncorrelated. We then moved on to the case of infiltration in statistically anisotropic soils. An unusually large single realization of strip source infiltration was simulated (300,000 node grid), under conditions similar to an on-going field experiment. Both the slope and the saturated value of the conductivity curve were taken random. Thus, a different conductivity curve was generated at each node of the grid. A detailed inspection of the pressure head field, sampled along transects and slices during 10 days of infiltration and 10 days of natural drainage, seemed to confirm some of the predictions of the linearized spectral theory (enhanced lateral spreading and pressure dependent anisotropy). Finally, we developed a more quantitative analysis for another large simulation in the case of steady "rainfall" infiltration on the same random soil realization (300,000 nodes). Spatial averaging estimates of pressure variability and unsaturated effective conductivity appeared fairly close to linearized spectral solutions. The chapter ends with a "summary and discussion" section, including a discussion of the current limitations and

future prospects of numerical and analytical approaches in the case of complex nonlinear flow systems.

This ends our description of the contents of chapters 3, 4, 5, 6, and 7. In summary, this dissertation was organized in three main topics: analytical/spectral approaches (Chapters 3 and 4), numerical analysis (Chapter 5), and statistical/physical interpretation of numerical simulations of random flow problems (Chapters 6 and 7). The conclusive *Chapter* 8 focuses on the implications of our findings for practical subsurface flow and contamination problems, and discusses some of the contributions of this work towards our conceptual understanding of stochastic flow and mass transport in heterogeneous porous media.

CHAPTER 2: REVIEW OF STOCHASTIC APPROACHES TO SUBSURFACE FLOW

2.1 Overview of Past and Current Approaches to Field Problems

2.1.1 Empirical models:

In view of the complexity of subsurface flow systems, most predictions were and still are based on empirical model calibration. Practitioners in the field of subsurface hydrology use numerical models based on local mass conservation, Darcy's law, and Fick's law. These are distributed-parameter models, usually two or three dimensional. Most frequently, natural variability is partially taken into account by dividing the flow domain into a few subdomains, with different conductivities. Layering within each block is also implicitly taken into account by specifying anisotropic conductivities, with a larger conductivity in the direction parallel to natural stratification, most often horizontal. The hydraulic and dispersive properties to be used in the model are further adjusted for a best fit between numerical and measured values (heads, concentrations). This calibration process has proved most expedient for addressing specific problems, but rather limited in scope. The typical situation is that accurate answers are obtained only for the particular conditions under which the model was calibrated.

Predictions made for hypothetical environmental conditions (different from those prevailing at the time of calibration) appear to be of limited value.

The limitations of current engineering practice are even more drastic concerning solute transport predictions. Dispersion coefficients obtained by way of model calibration appear much greater than those measured on small samples in the laboratory. Furthermore, they are found to increase as the contaminant spreads and more concentration data are made available for new calibrations. This inadequacy may be due to the fact that the velocity field predicted from the flow model, is much smoother than indicated by field observations (conductivity variability). Ignoring the small scale fluctuations of the flow field leads to inadequate prediction of the mechanical dispersion of convected species. Overall, it would seem that empirical calibration of these models does not allow for reliable predictions of contaminant migration over large time and length scales.

These remarks apply as well to black-box or zero-dimensional models. Jury et. al. (1982, 1986) proposed this type of model for the transport of solute in unsaturated soils. Briefly, their model isolates a soil unit (the black-box) and seeks to characterize a transfer-function that relates input and

output for that particular unit. This is analogous to the "Unit Hydrograph" method used in surface hydrology. Unfortunately, the data presented by Jury et al. (1982) indicate that the transfer functions calibrated for certain conditions do not extrapolate to other conditions, time scales, and length scales (see discussion in Gelhar et al. 1984).

2.1.2 Probabilistic models without spatial correlation:

There have been also a number of attempts at modelling unsaturated flow and transport by using the idea of independent soil columns. These approaches can be viewed as distributed-parameter models in one dimension. They take into account the horizontal variability of soil properties through a few random parameters, such as a random scaling factor in Warrick and Amoozegar (1979), Sharma et al. (1980), and Vauclin (1982); or the random saturated conductivity in the work by Dagan and Bresler (1983). These authors all assumed in effect that horizontal fluctuations of velocity were unimportant (one-dimensionality). The soil properties were assumed uncorrelated in the horizontal (statistically independent columns) but perfectly correlated in the vertical (homogeneous soil columns). However, other results obtained for fully three dimensional random properties indicate that *dimensionality* and *correlation scales* have a crucial influence on the overall

features of the flow field. The reduced dimensionality of the models mentioned above implies a smaller degree of freedom for the movement of fluid particles, whereas natural heterogeneities create fluid pathways that are inherently three-dimensional. In addition the assumption of statistically independent soil columns does not seem to make sense for columns of diameter smaller than the observed correlation scales. In fact, because of these simplifying assumptions the adequate inputs to be used in such models are difficult to evaluate. What are the average or effective parameters for each column? How are they related to small scale measurements?

2.1.3 Stochastic models with spatially correlated fields:

In view of the difficulties just mentioned, approaches based on the theory of random functions have been developed in an attempt to capture the essential features of flow and transport processes in heterogeneous media. In these stochastic approaches, local properties such as the hydraulic conductivity, or the storage coefficient are viewed as homogeneous random functions of space, with translation-invariant means, variances, and correlation functions, determined from field data. In actual practice, the data need only be approximately homogeneous. For instance, Ababou et al. (1985) argue that a statistically meaningful identification of a log-conductivity field must

satisfy a requirement of the type $\lambda << L << \mathcal{L}$, where λ is the correlation length or fluctuation scale, L is the size of the measurement network, and $\mathcal{L}$ is some scale of inhomogeneity (for instance $\mathcal{L} = d\ell nK_G/dx$ characterizes the length scale of inhomogeneity in the mean). In this framework, the governing equations (local mass conservation, Darcy's law, Fick's law) have stochastic solutions which can be characterized, in principle, in terms of their ensemble statistics (e.g. first and second order moments of heads, velocities and concentrations).

The infinite-domain spectral theory developed by Gelhar and others aims specifically at obtaining a large-scale characterization of spatially variable flow and concentration fields. By using ergodic arguments, they assume the equivalence between ensemble expectations and spatial averages. Their final results include close-form expressions for the head variance, velocity variance, effective conductivity and macrodispersion in a variety of situations (Bakr et al. 1978; Gelhar and Axness 1983; Gelhar 1984, 1986, and 1987). The "effective" transport properties were defined in connection with the "large scale" Darcy and Fick laws, relating mean fluxes to mean gradients (similar to the Onsager relations used in thermodynamics). The validity of these phenomenological equations have been traditionally investigated at the small scale only. What is emphasized here is that these equations may be extrapolated to

describe large scale phenomena, but with different coefficients. The effective conductivity and macrodispersion coefficients obtained from the spectral theory directly incorporate the spatial structure of the observed heterogeneities, in terms of the variance and correlation structure of the local hydraulic conductivities.

Other methods for solving stochastic flow and transport equations include the approximate Green's function method (Dagan, 1982), the direct numerical solution of approximate equations for first and second order moments (Townley, 1983; McLaughlin, 1985), and direct Monte-Carlo simulations (Freeze, 1975, Smith and Freeze 1979, Delhomme 1979, Ma et al. 1987). These methods apply in particular to the case of flow in bounded domains, and in the presence of local sources or sinks such as pumping wells. In the latter case, the ensemble moments of the stochastic flow field are to be interpreted in a Bayesian framework, i.e., they represent the uncertainty among many possible realizations or locations, rather than the large scale spatial variability of a single flow system (variance of heads near a pumping well at an unspecified location). This view was sometimes implicitly adopted, e.g. in the "well problem" treated by Dagan (1982). The case of semi-infinite aquifers was tackled by Naff and Vecchia (1986) through a combination of Green's function and spectral representation.

Some authors used Monte-Carlo simulations for checking analytical or other solutions. In most cases, comparisons were limited to fairly small conductivity variance (Freeze, 1975; McLaughlin, 1985). Indeed the number of realizations needed to obtain accurate answers could be prohibitive for the large conductivity variances observed in certain field sites, requiring the numerical solution of perhaps 10,000 or more flow problems in three dimensions (see discussion in McLaughlin, 1985). On the other hand, there are also questions about the range of validity of the perturbation-based solution methods just discussed (Gelhar, 1984; Dagan, 1982; McLaughlin, 1985): these are all in some sense "first order approximations", which are only valid asymptotically as the input variability becomes small. Even approximate indications on the robustness of these approximations would be extremely valuable, since exact solutions are not known.

2.2 The Single-Realization Approach:

2.2.1 Objectives and method:

In this section, we define briefly the single-realization approach which will be used in conjunction with a numerical solution method (Chapter 5) to obtain

representative solutions of stochastic flow problems (Chapters 6 and 7). It is relevant here to discuss in more rigorous terms what a "representative" solution means. The idea of the single-realization approach is that the statistical properties of the flow field can be evaluated by computing the spatial moments of one large realization of the flow, rather than the ensemble moments across a large number of realizations. The single-realization and ensemble approaches should be conceptually equivalent if the single solution is obtained on a sufficiently large domain, *and* if the ergodic hypothesis holds, as assumed in the spectral theory of Gelhar and others (see Chapter 3).

It may be useful here to briefly examine how large must "large" be in order to ensure the equivalence of ensemble and spatial moments of three-dimensional flow fields (e.g. hydraulic head or groundwater velocity vector). Consider a single finite realization of an ergodic flow field $Y(\underline{x})$ in three-dimensional space. Note that we do not question here the ergodicity hypothesis. Let us now focus on the behavior of spatial sample moments, say mean and covariance, as the size L_i of the three-dimensional domain varies. If λ_i is the correlation length of $Y(\underline{x})$ along x_i, and if $Y(\underline{x})$ is indeed statistically homogeneous along all three directions (i=1,2,3), it seems reasonable to define the sample size, or "equivalent number of independent samples", as:

$$N_Y = \frac{L_1 L_2 L_3}{\lambda_1 \lambda_2 \lambda_3} \tag{2.1}$$

Thus, as in the case of Monte-Carlo simulations, the classical theory of sampling errors can be used to determine the uncertainty on the computed spatial moments due to insufficient sample size. The relative errors on the mean and standard deviations $\overline{Y}$ and σ_Y are defined as:

$$\epsilon(\overline{Y}) = \frac{\sqrt{Var(\overline{Y})}}{\sigma_Y}$$
$$\epsilon(\sigma_Y) = \frac{\sqrt{Var(\sigma_Y)}}{\sigma_Y} \tag{2.2}$$

According to classical results of sampling theory (for a population of independent normal random variables) these quantities are both proportional to the inverse square-root of the equivalent sample size (see for instance Kendall and Stuart, 1977):

$$\epsilon(\overline{Y}) \simeq \frac{1}{\sqrt{N_Y}}$$
$$\epsilon(\sigma_Y) \simeq \frac{1}{\sqrt{2N_Y}} \tag{2.3}$$

However, the error on the covariance function $R_{YY}(\underline{\xi})$ will increase with lag-distance. To evaluate this effect in a very approximate manner, consider how the available number of samples decreases with lag-distance along, say, the x_1-axis:

$$\underline{\xi} = (\xi_1, 0, 0)$$

$$N_Y(\xi_1) = \frac{(L_1 - \xi_1)L_2L_3}{\lambda_1 \lambda_2 \lambda_3}$$

For an isotropic field ($\lambda_1=\lambda_2=\lambda_3$) and a cubic domain ($L_1=L_2=L_3$), the available number of samples that can be used to compute the covariance function at lag-distance $\underline{\xi}$ is then simply:

$$N_Y(\underline{\xi}) = \left(\frac{L}{\lambda} - \frac{\underline{\xi}}{\lambda}\right) \cdot \left(\frac{L}{\lambda}\right)^2 \qquad (2.4)$$

Now, we expect that the relative error on the covariance function:

$$\epsilon(R_{YY}) = \frac{\sqrt{\mathrm{Var}(R_{YY})}}{R_{YY}} \qquad (2.5)$$

will behave like:

$$\epsilon(R_{YY}(\underline{\xi})) \simeq \frac{1}{\sqrt{N_Y(\underline{\xi})}} \qquad (2.6)$$

Although these estimations of the uncertainty of sample

statistics are very approximate, they should suffice for our purpose here. More precise statements on sample statistics can be found in Appendix 2A, which will be used later in this work.

Let us now illustrate these approximate relations for a simple example: a cubic domain of size 5 correlation scales in each direction. The relative errors on the samples mean and standard deviation are reasonably low in this case (respectively 9% and 6%). However, the relative error on the sample covariance function increases with lag-distance: 10% at $\xi = \lambda$, 14% at $\xi = 3\lambda$, and over 20% for $\xi \geq 4\lambda$. Obviously, to obtain a reliable evaluation of the *spatial structure* of the flow field from a single realization, the flow domain must be taken much larger than the largest correlation scales of the variables describing the flow. In addition, the effects of artificial boundary conditions, anisotropic behavior, and intrinsic inhomogeneity of the flow along certain directions (e.g. parallel to flow) may lead to revise the size requirement up. This will be investigated more specifically in Chapter 6, using some of the sample statistics developed in Appendix 2A.

The requirement on the size of the single-realization flow problem (to ensure the equivalence of spatial and ensemble moments) is not unlike the Monte-Carlo method's requirement of a

large number of realizations in ensemble space. However, there is an essential difference, that has led us to favor the single-realization approach over Monte-Carlo. In the single-realization method, the spatial variability of a more or less homogeneous random phenomenon is studied at the large scale, and advantage is taken of the large size of the domain to sample the variables of interest over many fluctuation scales in physical space. In the Monte-Carlo method on the other hand, the large number of realizations required to obtain reliable ensemble moments is, in practice, incompatible with the study of *large scale flow or transport phenomena* by numerical methods. The Monte-Carlo method seems best adapted to the study of localized phenomena affected by uncertainty (uncertain drawdown near a pumping well, depending on its location in a heterogeneous aquifer). Note that our arguments are based on a two-sided interpretation of the effects of spatial variability, depending on the nature of the hydrologic problem and on the size of the domain of interest. We will touch upon this subject again in Chapter 4 (Section 4.4) with the idea of "spectral conditioning".

Furthermore, in cases where the ergodic hypothesis does not hold or is only approximately satisfied, the single realization approach can still be viewed as a *direct simulation* of plausible field conditions, where only one spatial realization of a particular heterogeneous flow system is actually available

for observation (over a finite domain). In contrast with the spectral theory, the numerical single realization approach does not postulate a priori the statistical homogeneity and ergodicity of the solution (although the assumption of a homogeneous/ergodic random porous medium is still retained). This brings the single realization approach closer to a realistic representation of natural field conditions. However, the statistical interpretation of just one spatial replica of a flow field may not make sense in situations where the solution appears to be strongly inhomogeneous, as would be the case for instance in the presence of some local source (single pumping well) or some inherent inhomogeneity at the domain-scale (geologic structure). A difficulty of this nature will be encountered for instance in the single-realization simulations of transient infiltration from a strip source (Chapter 7). In this case, the random flow solution is inherently inhomogeneous, especially at early times where there is a relatively well defined wetting front separating a wet and dry region. Nevertheless, even in the case of a moving front, the single-realization approach is still interesting because it provides, by *direct simulation*, a detailed picture of *one possible flow system* as may occur in natural conditions.

In summary, the numerical single realization approach of naturally heterogeneous flow systems seems to be a natural way to obtain, by direct simulation of the flow field, a realistic

picture of the effects of (random) spatial variability on the flow pattern, with a degree of detail that cannot be achieved by direct subsurface measurements in the field. More importantly, the simulated flow field can be statistically analyzed in space and compared to the ensemble results of the spectral theory, in cases where there is enough statistical homogeneity for such a comparison to make sense. This method of analysis will be applied systematically to the numerical solutions obtained for steady state groundwater flow (Chapter 6), and also for a problem of steady state infiltration (Chapter 7). As mentioned above, the case of transient infiltration from a local source (Chapter 7) produces inherent inhomogeneities in the flow pattern, and, precisely for this reason, a statistical analysis will not be attempted there. Nevertheless, some qualitative comparisons with the results of the linearized spectral theory will be developed based on visual observations of the spatial pattern and evolution of the wet zone.

2.2.2 Generation of Random Fields by the Turning Band Method:

To be useful, the numerical single-realization method also requires the accurate generation of random field hydraulic properties on a relatively fine three-dimensional grid. The mesh size of the grid must be small enough that the statistical properties of inputs (e.g. log-conductivity) and outputs (e.g.

heads and water velocities) be preserved in the discrete representation. Without going into details, it seems reasonable at first sight to require that:

$$\Delta x/\lambda \ll 1 \qquad (2.7)$$

where λ is a typical fluctuation scale of the flow. This requirement is similar to the sampling theorem in signal theory: a temporal signal of period T must be sampled at time intervals $\Delta\tau \ll T/2$ in order to avoid aliasing. In the present case, the fluctuation scale λ plays the role of the half-period T/2.

A more quantitative interpretation of the resolution constraint (2.7) can be developed by evaluating the behavior of the local integral of the random field of interest over the discrete cell of size Δx. The resolution of the grid must be fine enough that the statistical properties of the locally integrated field $\overline{Y}(\underline{x})$ be close to the random field $Y(\underline{x})$ defined in continuous space. For a given value of $\Delta x/\lambda$, there is a reduction in variance and an increase in the correlation scale of $\overline{Y}(\underline{x})$ compared to the point process $Y(\underline{x})$. For instance, in the simple case of a one-dimensional process $Y(x)$ with exponential covariance function, it is not difficult to see that the statistics $(\sigma^2_{\Delta x}, \lambda_{\Delta x})$ of the locally integrated process:

$$\bar{Y}(x) = \frac{1}{\Delta x}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} Y(x')dx' \qquad (2.8)$$

are related to the (σ^2,λ) statistics of the point process Y(x) by:

$$\frac{\sigma^2_{\Delta x}}{\sigma^2} = 2\,\frac{\lambda}{\Delta x}\,\left\{1 - \frac{\lambda}{\Delta x}\,(1 - e^{-\Delta x/\lambda})\right\}$$

$$\frac{\lambda_{\Delta x}}{\lambda} = \frac{\sigma^2}{\sigma^2_{\Delta x}}. \qquad (2.9)$$

Thus, the distortion in the variance and correlation scale of Y(x) upon local integration over a mesh of size $\Delta x = \lambda/4$ is about 10%. This may give a rough idea of the quantitative meaning of the inequality constraint (2.7): it seems reasonable to accept a grid resolution equal to a fraction of unity, i.e not necessarily "much smaller than" unity. A more precise analysis of the grid resolution requirement will be developed in Chapter 5 in connection with the numerical issues pertaining to the discretized solution of stochastic flow equations.

At any rate, once the mesh size has been chosen, it is also important to be able to generate a *representative* realization of statistically homogeneous random field hydraulic

properties with specified single-point and n-point moments. We used for this purpose a three-dimensional version of the "turning band" random field generator, recently developed by Tompson, Ababou and Gelhar (1987). The idea of the turning band method was brought up by Matheron (1973) for multidimensional homogeneous and isotropic random fields. Its practical implementation was developed by Mantoglou and Wilson (1982), particularly for two-dimensional random fields. The case of three-dimensional isotropic *and* anisotropic fields is treated at length in Tompson et al. (1987), including a number of numerical experiments. Ababou (1986) also discusses in an unpublished report some possible extensions of the method to treat the case of *self-similar* random fields, and Mantoglou (1987) elaborates on the case of statistically homogeneous *vector* fields.

Let us *briefly* describe the principle of the turning band method in the case of a statistically homogeneous random field $f(\underline{x})$ having a zero-mean Gaussian distribution, and an isotropic or ellipsoidal covariance function in three-dimensional space. The method relies on the representation of $f(\underline{x})$ in Fourier space (see Chapter 3):

$$f(\underline{x}) = \int_{-\infty}^{+\infty} e^{i\underline{k}\underline{x}} \, dZ_f(\underline{k})$$

$$R_{ff}(\underline{\xi}) = \int_{-\infty}^{+\infty} e^{i\underline{k}\underline{x}} \, S_{ff}(\underline{k}) d\underline{k} \qquad (2.10)$$

where $R_{ff}(\underline{\xi})$ is the covariance function, $dZ_f(\underline{k})$ the complex Fourier-Stieltjes increment, and $S_{ff}(\underline{k})$ the spectral density of $f(\underline{x})$ defined by:

$$S_{ff}(\underline{k}) \cdot d\underline{k} = \langle |dZ_f(\underline{k})|^2 \rangle \tag{2.11}$$

In particular, in the 3D isotropic case, the covariance function in (2.10) can be expressed directly in terms of a radial spectrum as follows (Adler, 1981):

$$\begin{aligned} R_{ff}(\xi) &= 2 \int_0^\infty \frac{\sin(k\xi)}{k\xi} E(k)\, dkk \\ E(k) &= \int_{\mathcal{S}(k)} S(\underline{k}) d\sigma(\underline{k}) = 4\pi k^2 S(k) \end{aligned} \tag{2.12}$$

where the integral defining the radial spectrum E(k) is taken over the sphere of radius k in $\mathbb{R}^3$. Note that k and ξ are the *radial* wavenumber and separation distance, respectively.

The idea of the turning band method is that it is possible to find a one-dimensional process $f_1(x)$ generated independently along lines having many different orientations in space:

$$f(\underline{x}) = \lim_{L\to\infty} \frac{1}{\sqrt{L}} \sum_{i=1}^{L} f_1^{(i)}(\underline{x} \cdot \underline{u}_i) \qquad (2.13)$$

Each line (i) is defined by its direction vector $\underline{u}_i$, and $f_i^{(i)}(x)$ is the sample function of process f_i generated along line (i) as illustrated in Figure (2.1). The 1D sample functions are generated independently of each other, and the line direction vectors $\underline{u}_i$ are drawn from a random unit vector having a uniform distribution on the sphere (see Tompson et al. 1987). In these conditions, it is not difficult to see that the covariance function of the 3D field defined by (2.13) takes the form:

$$\boxed{R_{ff}(\underline{\xi}) = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} R_i(\underline{\xi}\cdot\underline{u}_i)} \qquad (2.14)$$

where $R_i(\underline{\xi})$ is the covariance function of the 1D process $f_i(x)$. With additional assumptions of homogeneity and ergodicity, on which we do not elaborate here, the sum in (2.14) can be replaced by an ensemble average over all possible realizations of the *random* direction vector $\underline{u}_i$ as follows:

$$R_{ff}(\underline{\xi}) = \langle R_i(\underline{\xi}\cdot\underline{u}) \rangle \qquad (2.15)$$

Finally, with a uniform distribution of $\underline{u}$ on the sphere, equation

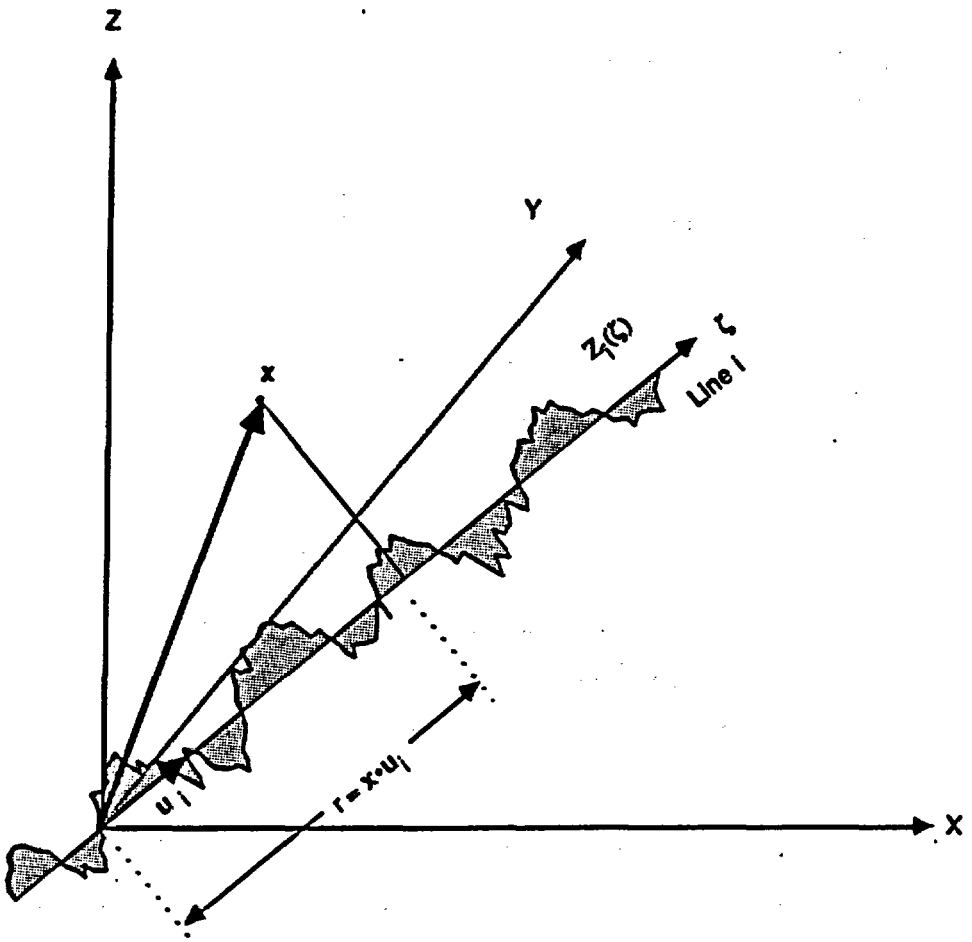


Figure 2.1: Turning band method: projection of the i-th line process $f_i^{(i)}(x)$ onto an arbitrary point $\underline{x}$ in three-dimensional space (from Tompson, Ababou, Gelhar, 1987).

(2.15) leads to an explicit relation between the covariance function of the line process and the desired covariance function of the isotropic random field to be generated:

$$R_{ff}(\xi) = \frac{1}{4\pi} \int_0^{2\pi} d\theta \int_0^{\pi} R_1 (\xi \cos\phi) \sin\phi \, d\phi$$

$$= \frac{1}{\xi} \cdot \int_0^{\xi} R_1(s) ds$$

whence:

$$\boxed{R_1(\xi) = \frac{d}{d\xi} (\xi \, R_{ff}(\xi))} \qquad (2.16)$$

Equation (2.16) gives explicitly the covariance of the line-process such that the projection operator (2.13) yields the desired 3D isotropic covariance $R_{ff}(\xi)$, asymptotically as the number of lines L goes to infinity. The method used in the turning band algorithm to generate each line process is the spectral decomposition method of Shinozuka and Jan (1972). Details are given in Tompson et al. (1987). In actual practice, the number of lines need not be very large to obtain relatively accurate single-realizations as far as first and second moments are concerned. In particular, it seems that the number of lines

L required to be consistent with the resolution of the grid may only grow like some fractional power of the total number of nodes of the grid. This empirical observation seems to be confirmed by the results of Tompson et al. 1987, and those obtained in this work: see Chapter 6 (Section 6.3) for a realization of the log-conductivity field on a 1 million node grid using the turning band generator with 1000 lines.

Finally, the case of *ellipsoidal* anisotropic random fields does not require any modification of the above method. For any isotropic field with covariance function $R_{ff}(\xi)$ and fluctuation scale λ, an ellipsoidal field can be constructed by rescaling the three coordinates as follows:

$$\xi'_i = \frac{\lambda}{\lambda_i} \xi_i \qquad (2.17)$$

Thus, the new ellipsoidal covariance function $R'_{ff}(\xi)$ with fluctuation scales (λ_i) is simply given by:

$$R'_{ff}(\xi) = R_{ff}(\sqrt{\Sigma\ (\frac{\lambda}{\lambda_i} \xi_i)^2}) \qquad (2.18)$$

This simple relation was used in Chapters 6 and 7 to obtain single-realizations of statistically anisotropic (*ellipsoidal*) random hydraulic properties. For other applications such as

those involving space-time processes, more general classes of anisotropic random fields may be needed. Tompson et al. (1987) discussed this case as well, and Sivapalan and Wood (1986) used the 3D turning band generator to simulate anisotropic space-time rainfall intensity fields.

2.3 Brief Survey of Field Data

2.3.1 Hydraulic properties of heterogeneous aquifers

In the case of groundwater flow, the local hydraulic properties of interest are the saturated conductivity K (m/s) and the specific storativity S_s (m^{-1}), or their two-dimensional equivalents, the transmissivity T (m^2/s) and the storage coefficient or specific yield S_y (dimensionless).

The spatial variability of hydraulic conductivity and transmissivity have been the object of intensive experimental studies in the recent past, in an effort to characterize their variability in a statistical rather than purely descriptive fashion. Gelhar (1986) reviews some of the available field data. Most experimental studies of aquifer variability actually focused on the horizontal variability of the transmissivity or of some depth-averaged conductivity determined from well pumping tests (Delhomme 1979, Binsariti 1980, Devary and Doctor 1982, Hoeksema

and Kitanidis 1985). Their results could be summarized as follows: the standard deviation of log-transmissivity (ℓnT) ranged between 0.6 (sandstone aquifer of Hoeksema and Kitanidis) to 2.3 (limestone aquifer of Delhomme), with correlation ranges on the order of 1 km to several tens of kilometers. The domain sizes were on the order of 5 km to several hundred kilometers.

Other studies concerned the local hydraulic conductivities obtained from small scale measurements, usually along vertical boreholes (Bakr 1976, Hufschmied 1985, Sudicky 1986). In these three studies of aquifer variability, the standard deviation of log-conductivity (ℓnK) ranged from 0.6 to 2.2, with correlation ranges on the order of 0.1 to 1.0 meters in the vertical. The length of the vertical transects (boreholes) was 20 to 100 meters. The studies of Hufschmied (1985) and Sudicky (1986) are particularly remarkable because these authors actually determined, by different methods, the *three-dimensional* variability of the log-conductivity field. Hufschmied used a relatively sophisticated flowmeter measurement of conductivities in 16 wells, for twenty 1 meter thick layers in each well. The statistics of the point process $\ell nK(\underline{x})$ in the vertical were then obtained by using statistical identities relating the point process to the local average process. Information obtained among different wells was used indirectly to infer the horizontal covariance structure of $\ell nK(\underline{x})$. On the other hand, Sudicky

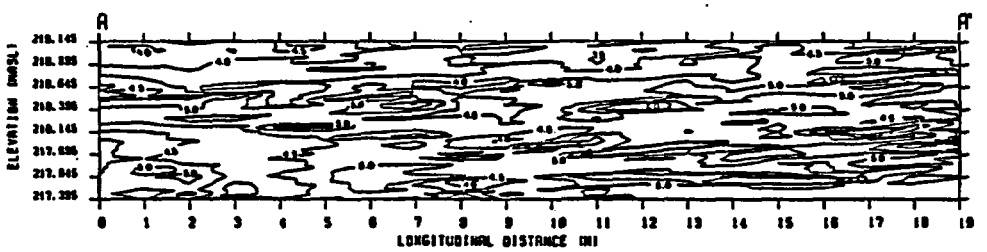
(1986) measured the conductivities from small core samples along a number of vertical boreholes organized along two perpendicular planes. This particular choice allowed them to estimate the covariance structure of ℓnK in three-dimensional space.

The statistical results obtained by Hufschmied (1985) and Sudicky (1986) are summarized in Table 2.1. In addition, Figure (2.2) (top) shows the log-conductivity contours obtained by Sudicky (1986) along the vertical plane aligned with the natural hydraulic gradient at the Borden site. For comparison, the bottom part of Figure (2.2) displays an artificially generated anisotropic random field obtained by the turning band method described previously. Note that both Hufschmied and Sudicky's data show a significant statistical anisotropy between the horizontal and vertical directions.

The data reviewed above indicate that the conductivities and transmissivities can vary over several orders of magnitudes in natural formations. In addition, the work of Hufschmied (1985) and Sudicky (1986) clearly shows that the conductivities follow a log-normal probability distribution more closely than a normal distribution. Finally, it is also clear that in most cases, there is a relatively strong anisotropy in the vertical/horizontal correlation scales of the log-conductivity (Table 2.1). However, it is also possible that

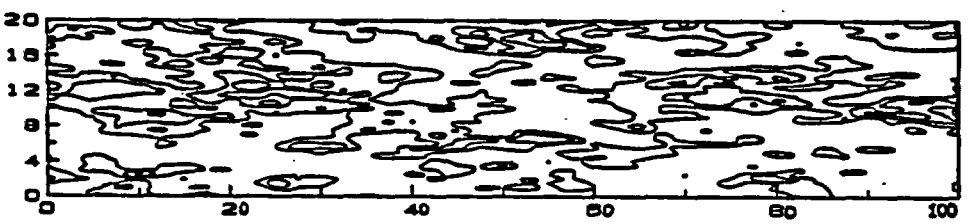
TABLE 2.1: STATISTICAL PROPERTIES OF MEASURED THREE-DIMENSIONAL LOG-CONDUCTIVITIES (ℓnK) IN SATURATED FORMATIONS, FROM HUFSCHMIED (1985) AND SUDICKY (1986)

	Hufschmied (1985)	Sudicky (1986)
Site	Aeflingen, Switzerland	Borden site, Ontario
Formation	20 m thick, sand and gravel	Outwash sand
Covariance	Anisotropic exponential	Anisotropic exponential
K_G	($6\ 10^{-3}$ m/s)	$7.17\ 10^{-5}$ m/s
σ_f	1.92	0.54 - 0.62
λ_1	15 - 20 m	2.8 m
λ_2	15 - 20 m	2.8 m
λ_3	0.5 m	0.12 m



Location of measurements and distribution of -ln(K) along A-A' (contour interval = 0.5; vertical exaggeration = 2; K < 10^{-3} cm/s in stippled zones).

(a): Measured



(b): Simulated by the Turning Band Method

Figure 2.2 Log-conductivity contours in a vertical plane. The upper part is reproduced from Sudicky, 1986 (measured at the Borden tracer site). The lower part of the figure was obtained by simulation, using the Turning Band Method with an anisotropic spectrum.

the anisotropy observed by these authors be due in part (or depend in part) on the discrepancy between the horizontal and vertical scales of the domain under investigation (12 m x 3 m in the study of Sudicky, 1986). Indeed, some of the other studies mentioned above suggest that the supposedly stationary correlation structure of porous formations may not be well defined for certain sites and/or for certain domain sizes. This is suggested for instance by the statistical analysis of Hoeksema and Kitanidis (1985). These authors found in a number of cases that the correlation scale was either small (on the order of measurement spacing) or large (on the order of domain size). In addition, both the correlation structure and the degree of variability may be strongly influenced by certain subjective choices, such as empirical detrending (an example can be found in Devary and Doctor, 1982). These difficulties show that the measured statistical properties of supposedly homogeneous conductivity fields may be in fact scale-dependent in certain situations. In our view, this is not necessarily an obstacle to the application of stochastic concepts as long as the fact is recognized, and that adequate methods be designed to "pass" from one scale of analysis to another. The need for variable scale analysis may arise for predicting contaminant migration over large time scales. This particular question will be examined in Chapter 4 (Section 4.4).

For completeness, let us mention the papers by Clifton and Neuman (1982) and Hoeksema and Kitanidis (1985) concerning the spatial variability of the specific storativity of aquifers. The first authors found a positive correlation between log-transmissivity and log-capacity, with a slope close to unity. The specific capacity does not play a role in the present work, since only steady state problems will be considered in our analyzes of stochastic groundwater flow.

2.3.2 Constitutive relations of heterogeneous soils:

In this work, we will be concerned with both transient and steady state flow problems in unsaturated soils (Chapter 7). To describe these phenomena, two constitutive relations need to be determined:

- the water retention curve $\theta(h)$ relating volumetric moisture content to pressure head.
- the unsaturated conductivity curve $K(h)$ relating hydraulic conductivity to pressure head.

Our particular choice of functional relationships for $\theta(h)$ and $K(h)$ will be defined in Chapter 5 (Section 5.1). However, experimentators have used a variety of functional shapes to describe the $\theta(h)$ and $K(h)$ relationship in view of analyzing

their spatial variability. This makes it quite difficult to synthesize their findings in a compact form.. Nevertheless, we have attempted to summarize in a table some of the most significant results on natural soil variability in the literature.

Table 2.2 gives information on the probability distribution function and the correlation structure of the hydraulic parameters intervening in the constitutive relationships $\theta(h)$ and $K(h)$. It should be mentioned that these results correspond to different scales of analysis and different schemes of data collection, so they are not always directly comparable. However, some common features seem to emerge as a whole. First of all, it should be emphasized that only one author (Russo, 1983) analyzed the correlation structure of the whole conductivity curve $K(h)$, including in particular the correlation structure of the saturated conductivity $K_s(x)$ as well as that of the shape parameter:

$$\alpha(x) = \left.\frac{\partial \ell n K}{\partial h}\right|_x$$

His results indicate that both α and K_s are log-normally distributed, and that the correlation range of $\ell n\alpha(x)$ is significantly larger than that of $\ell nK_s(x)$.

TABLE 2.2: SUMMARY OF SOME FIELD DATA ON THE SPATIAL VARIABILITY OF UNSATURATED SOILS FROM THE LITERATURE

Y	$\bar{Y}$	CV_Y	$\sigma_{\ell n\ Y}$	τ_Y	λ_Y	Ref. and Remarks
K_o (cm/h)	1.96	0.71	0.64	2.49	a = 38 m λ = 17m Spherical	Russo, 1983 Hamra Red Mediterranean $K(h) = K_o e^{-\alpha h}$
α (cm^{-1})	.0286	0.43	0.41	1.36	variogram a = 51 m λ = 72 m Exponent. variogram	A = 2 + 2.5 ha N = 25 + 6 plots 1 plot = 3m x 3m n = 5 depths a = range of variogram λ = integral corr. length
θ_s	.338	0.08	0.08	(0.23)	76m	Russo, Bresler 1981
θ_r	.030	0.30	0.29	(0.93)	39 m	Hamra Read Mediterranean $\frac{\theta-\theta r}{\theta_s-\theta_r} = (\frac{h_w}{h})^{\beta}$
K_o (cm/h)	22.0	0.41	0.39	(1.2)	34m	$\frac{K}{K_o} = (\frac{h_w}{h})^{\eta}$
β	1.160	0.68	0.62	2.35	23m	$\eta = 2\beta + 2$
h_w (cm)	-7.2	0.22	0.22	(0.7)	48m	A = 0.8 ha N = 30 locations n = 4 depths (depth z=0 shown here)
θ_s	.397	0.10	0.10	(0.29)	—	Nielsen et al., 1973
K_o (cm/h)	0.85	1.06	0.87	4.37	—	Panoche soil A = 150 ha N = 20 plots n = 6 depths (depth z = 30 cm shown here)

TABLE 2.2: SUMMARY OF SOME FIELD DATA ON THE SPATIAL VARIABILITY OF UNSATURATED SOILS FROM THE LITERATURE (CONTINUED)

Y	$\bar{Y}$	CV_Y	$\sigma_{\ell n\ Y}$	γ_Y	λ_Y	Ref. and Remarks
K_o (cm/h)						Scisson, Wierenga, 1981 $A \simeq 40m^2$ 3 infiltration rings: Diameter Number
(1)	0.27	0.70	0.63	2.45	0.13m	ϕ_1 = 5cm N_1 = 625
(2)	0.35	0.54	0.51	1.79	—	ϕ_2 = 25cm N_2 = 125
(3)	0.36	0.22	0.22	0.68	—	ϕ_3 = 127cm N_3 = 25
K_o (cm/h)	0.70	0.40	0.38	(1.26)	50 m (10-50m)	Vieira et al., 1981 A = 0.88 ha N = 1280 plots 160 plots/transect 55 transects
K_o (cm/h)	1.46	0.60	0.56	2.02	—	Sharma et al. 1980 Watershed A = 9.6 ha N = 26 plots
K_o (cm/h)	0.13	1.02	0.84	4.10	0-2 m	Luxmoore et al. 1981 A = 192 m^2 N = 48 plots

The log-normality of K_s, α, and other inherently positive unsaturated soil parameters, seems to be confirmed by other authors. For convenience, we have listed in Table 2.2 the skewness coefficient τ. When indicated in parenthesis, the skewness was computed by us based on the assumption of *log-normality* instead of the normal distribution assumed by the authors (references are given in the right-most column). Thus, in all cases, the skewness τ_Y was computed from the classical identity:

$$\tau_Y = 3\ CV_Y + (CV_Y)^3 \tag{2.19}$$

where CV_Y is the coefficient of variation of the log-normal random variable Y. In addition, we also used the following identities:

$$(CV_Y)^2 = \frac{\sigma_Y^2}{\overline{Y}^2} = \exp\ (\sigma_{\ell nY}^2)-1 \tag{2.20}$$

$$\overline{Y} = Y_G \exp\ (\sigma_{\ell nY}^2/2) \tag{2.21}$$

where Y_G is the geometric mean defined by:

$$Y_G = \exp\ (\overline{\ell nY}).$$

These relations (2.19 - 2.21) can be found for instance in Vanmarcke (1983). They were used to compute the ℓnY-statistics in Table 2.2, when not provided by the authors themselves. It appears that a number of authors adopted the normality assumption precisely in those cases where the variability was too mild to be able to distinguish a normal from a log-normal distribution. Overall, we conclude that all the parameters listed in the table were in fact more or less positively skewed, and are presumably better represented by log-normal distributions. Figure (2.3) illustrates the difference between a normal and a log-normal random function (artificially generated for illustration). Note the sharp maxima and smooth minima typical of a postively skewed process.

Another feature that emerges from Table 2.2 is that the unsaturated conductivity curve seems *in general* to have greater variability than the water retention curve (although this is not always verified for a given site). The coefficient of variation of K_s was larger than unity in four cases, whereas the coefficient of variation of parameters involved in the $\theta(h)$ relation was always below or at most equal to 0.6.

The data of Table 2.2 do not contain any information on the correlation structure of soil properties in the vertical

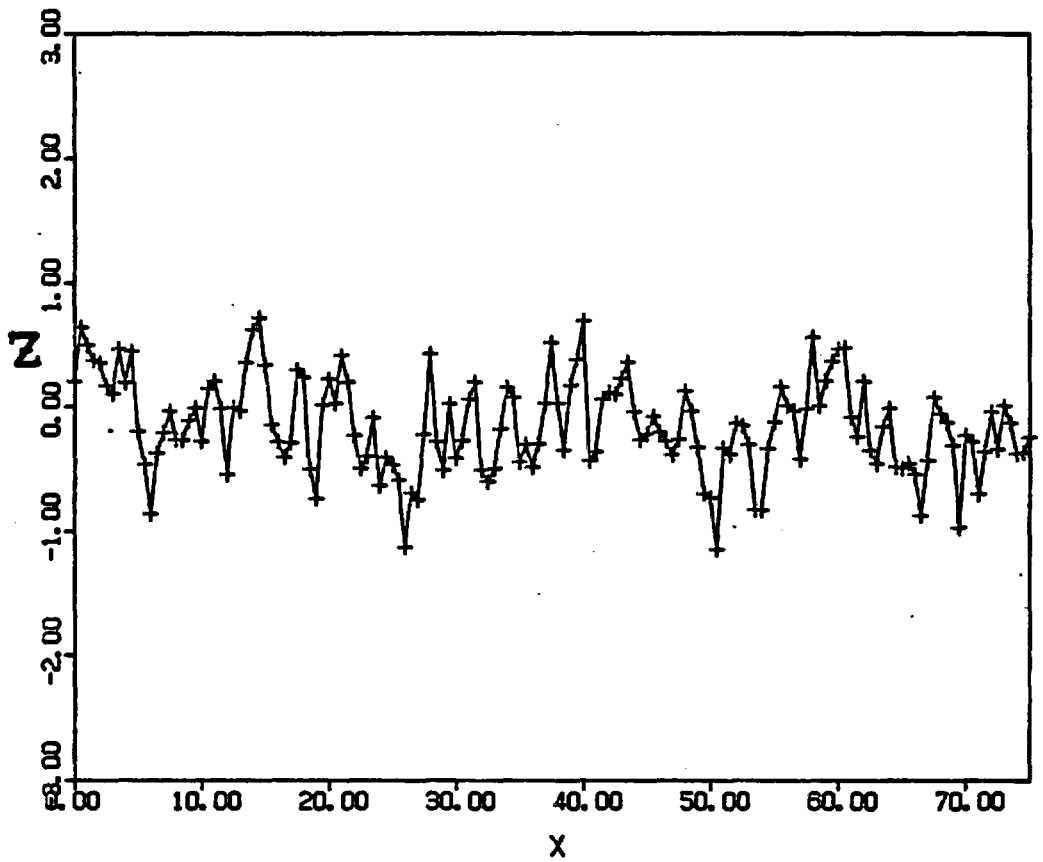


Figure 2.3(a): Normally distributed random function Z = ℓn Y ($-\infty$ < ℓn Y(x) < + ∞).

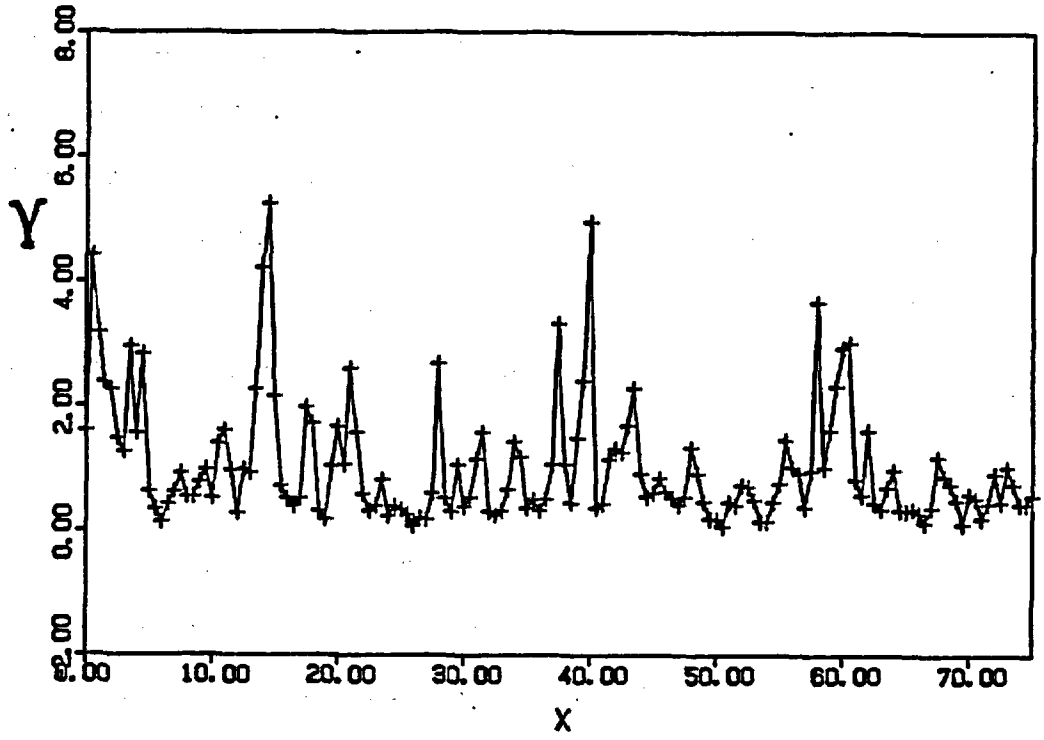


Figure 2.3(b) Log-normally distributed random function Y
$(0 < Y(x) < +\infty)$.

direction. The horizontal correlation scales range from 10 cm to a few tens of meters. These numbers should be taken "with a grain of salt", as the apparent correlation scales are most probably influenced by the density and size of the measurement networks. In addition, the results of Scisson and Wierenga (1981) indicate that the scale of the instrument (infiltration ring) has a definite effect on variability. Finally, it should also be noted that the correlation structures were only determined for relatively small domains, on the order of 1 hectar for Russo and Bresler 1981, Vieira et al. 1981, and Russo 1983. The largest domain of investigation was the 150 ha site of Nielsen et al. 1973; however these authors did not determine any correlation structure.

In order to illustrate more concretely the spatial variability of soil properties, we have reproduced in Figure 2.4 the unsaturated conductivity curves obtained by Nielsen et al. (1973) on the Panoche silty clay loam. This figure indicates that both the saturated conductivity and the slope of the $\ell nK(h)$ curve are fairly variable. The figure does not seem to indicate a strong degree of correlation between these two parameters, although some correlation could be expected on physical grounds: coarser soils are generally more permeable at saturation *and* have a steeper $\ell nK(h)$-slope (Ababou, 1981). The effect of correlation between K_s and the $\ell nK(h)$-slope will be investigated by way of

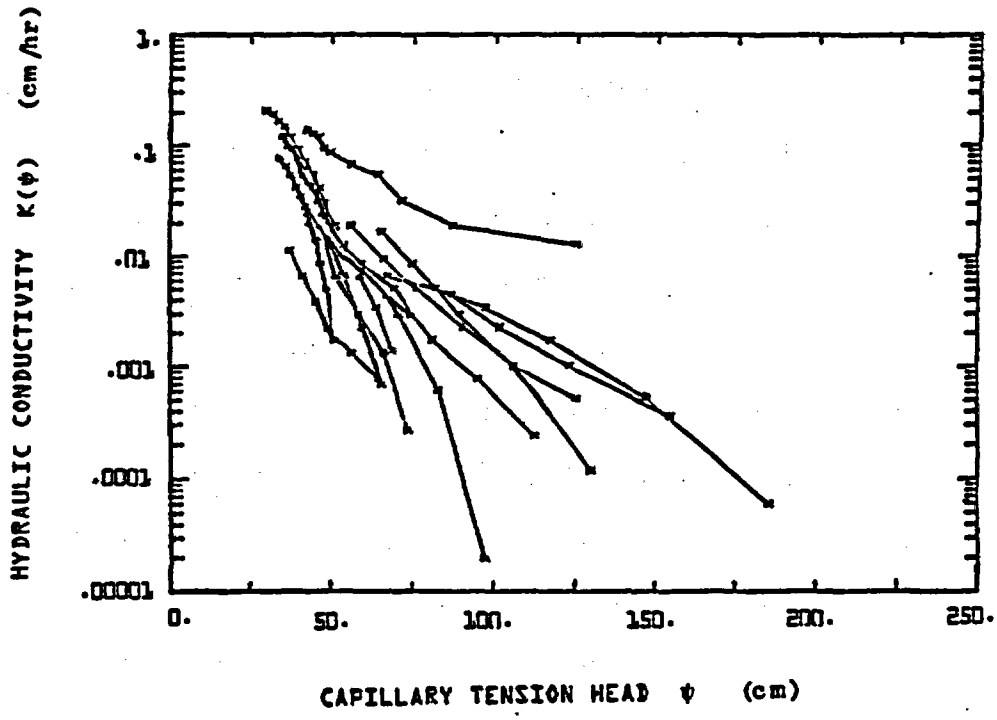


Figure 2.4: Unsaturated hydraulic conductivity versus capillary tension head for the Maddock sandy loam. Each curve corresponds to a different spatial location (from Yeh et al., 1982).

numerical experimentation in Chapter 7 (Section 7.2). Additional data and references concerning an on-going strip-source infiltration at the University of New Mexico at Las Cruces will be given in Chapter 7 (Section 7.3).

Our literature review, partially summarized in Table 2.2, indicated that available data are too scant to completely characterize the spatial variability of unsaturated soils hydraulic properties in terms of random field parameters. The variability of the unsaturated conductivity curve in particular seems to have important effects on the behavior of unsaturated flow (Mantoglou and Gelhar, 1987). Unfortunately, no comprehensive study of the three-dimensional spatial structure of $\ell n\ K(h)$ has yet been undertaken. However, some of the missing data could be inferred directly by correlating, for instance, the shape of the $\ell n\ K(h)$ curve to the value of the saturated conductivity for different types of soils. Other alternative approaches of this kind have had some success, for instance those relying on the "similar media" postulate. This latter approach reduces the characterization of spatial variability of $\theta(h)$ and $K(h)$ to just one spatially random scaling factor: cf. experimental studies of Warrick et al. (1977), Simmons et al. (1979), Russo and Bresler (1980), Vauclin et al. (1981), and Sharma et al. (1980). The field remains open: future research could focus on such correlations or similarity assumptions, as a

means of characterizing in a more comprehensive fashion the spatial variability of the nonlinear constitutive relations of unsaturated media.

At any rate, the complexity of unsaturated flow processes in heterogeneous formations should not be overlooked. Figure 2.5 shows the spatial structure and evolution of the wetted zone during an infiltration experiment in the Hanford sediments (from Routson et al., 1979). The wetted zone in this case is remarkably assymmetric and has pronounced lateral spreading. To explain these features statistically requires a model of three-dimensional variability that includes the effects of stratification, e.g., statistical anisotropy. Similar effects of assymmetry and spreading will be observed in the numerical simulations of Chapter 7 (Section 7.3) with fully three-dimensional and statistically anisotropic random conductivity curves.

(a) After 6 hours

(b) After 24 hours

Figure 2.5 Typical horizontal and vertical movement of liquids in Hanford formation sediments under partially saturated conditions. Taped area outlines position of water addition (from Routson et al., 1979).

CHAPTER 3: FIRST ORDER SPECTRAL SOLUTIONS FOR STOCHASTIC FLOW IN SATURATED MEDIA

3.1 Formal Solution of Spectral Perturbation Equations:

In this section, we develop approximate solutions for the stochastic equation of steady state flow in random saturated media, following the first order spectral theory previously developed by *Bakr et al. 1978*, and *Gelhar and Axness 1983*. The assumptions made at each step will be clearly stated, in anticipation of subsequent discussions on the approximate nature of perturbative spectral solutions.

Our point of departure is the partial differential equation governing flow in a saturated porous medium with spatially variable conductivity $K(\underline{x})$. This equation is obtained from the steady state mass conservation equation and the local Darcy equation, (Darcy, 1856) respectively:

$$\nabla \underline{Q} = 0. \tag{3.1}$$

$$\underline{Q} = -K(\underline{x}) \cdot \nabla \underline{H}. \tag{3.2}$$

Here, $\underline{Q}$ represents the specific discharge rate (Darcy velocity), and H the hydraulic head potential. Using (3.2) in (3.1) leads to the familiar groundwater flow equation:

$$\nabla(K(\underline{x})\cdot\underline{\nabla}H) = 0. \qquad (3.3)$$

Let us assume that the conductivity $K(\underline{x})$ in (3.3) is a second order stationary random function in 3D space, with a log-normal distribution. Therefore, the statistical properties of the *log-conductivity* $F(\underline{x})$ are entirely described by its first and second order moments, i.e., its mean $\overline{F}$ and covariance function R_{ff} as defined below:

$$\begin{cases} F(x) = \ell n\ K(\underline{x}) \\ \langle F(\underline{x})\rangle = \overline{F} = \ell n\ K_G \\ \langle (F((\underline{x})-\overline{F}).(F(\underline{x}')-\overline{F})\rangle = R_{ff}(\underline{\xi}), \qquad \underline{\xi} = \underline{x} - \underline{x}' \end{cases}$$

where K_G is the geometric mean conductivity, and $\underline{\xi}$ a separation vector. The flow equation (3.3) can be decomposed to be expressed in terms of F as follows:

$$\boxed{\nabla^2 H + \underline{\nabla}F(\underline{x})\cdot\underline{\nabla}H = 0.} \qquad (3.4)$$

Equation (3.4) will be used for the subsequent spectral perturbation analysis. Observe that (3.4) is a stochastic partial differential equation, due to the random character of the log-conductivity gradient vector $\underline{\nabla}F(\underline{x})$, even in the case of deterministic boundary conditions.

Let us now express (3.4) in terms of ensemble averages and perturbations. The perturbations of the random fields log-conductivity and hydraulic head are defined as:

$$f(\underline{x}) = f(\underline{x}) - \langle f(\underline{x})\rangle = \ell n\,[K(\underline{x})/K_G]$$
$$h(\underline{x}) = H(\underline{x}) - \langle H(\underline{x})\rangle$$

The governing flow equation can be separated into a mean equation and a perturbation equation, by substituting $H = \langle H\rangle + h$ and $F = \langle F\rangle + f$ in (3.4) and applying ensemble averaging operators. The mean equation obtains directly by ensemble averaging equation (3.4):

$$\nabla^2 \langle H\rangle + \nabla\langle f\rangle\cdot\nabla\langle H\rangle = -\langle \nabla f\cdot\nabla H\rangle$$

and the perturbation equation obtains by subtracting the above mean equation from (3.4):

$$\nabla^2 h + \underline{\nabla}\langle F\rangle\cdot\underline{\nabla} h + \underline{\nabla} f\cdot\underline{\nabla}\langle H\rangle = -\{\nabla f\cdot\nabla h - \langle f\cdot\nabla h\rangle\}.$$

Finally, using the assumption that the mean log-conductivity $\langle F\rangle$ is constant (K_G constant), we obtain:

$$\nabla^2\langle H\rangle = -\langle \underline{\nabla} f\cdot\underline{\nabla} h\rangle \qquad \textit{(mean eq.)} \qquad (3.5.a)$$
$$\nabla^2 h + \underline{\nabla} f(\underline{x})\cdot\underline{\nabla}\langle H\rangle = -\{\nabla f\cdot\nabla h - \langle \nabla f\cdot\nabla h\rangle\} \quad \textit{(perturbation eq.)} \qquad (3.5.b)$$

Observe that equations (3.5) form a system of two stochastic equations with three unknowns: $h(\underline{x})$, $\langle H(\underline{x})\rangle$, and the second order term $\langle \underline{\nabla} f \cdot \underline{\nabla} h \rangle$. A complete solution would involve solving an infinite hierarchy of equations governing higher order moments. This turns out to be an impossible task in the general case. Various expansion methods have been proposed in the literature in order to arrive at approximate close-form solutions (see review in Chapter 2). We develop below first order perturbation approximations following the work of Gelhar and others quoted above, although in a slightly different manner.

One way to obtain a first order solution is to expand the solution $H(\underline{x})$ in powers of σ_f, the "small parameter" in the expansion (the possibly divergent character of this expansion will be discussed in Section 4.1). Accordingly, let:

$$H = H_0 + \sigma H_1 + \sigma^2 H_2 + \cdots .$$

On the other hand, note that the random field $f(\underline{x})$ in (3.4) is proportional to $\sigma_f = \sigma$, its standard deviation. For a correct perturbative analysis, we must use instead a normalized random field $g(\underline{x})$ with unit variance and zero mean:

$$g(\underline{x}) = f(\underline{x})/\sigma \quad .$$

The conductivity is simply related to $g(\underline{x})$ by:

$$K(\underline{x}) = K_G \exp(\sigma g(\underline{x})).$$

Plugging $g(\underline{x})$ in equation (3.4) and assuming K_G constant gives:

$$\nabla^2 H + \sigma \nabla g \cdot \underline{\nabla} H = 0.$$

Plugging the expansion for H in this equation yields the infinite hierarchy of equations:

$$\begin{cases} \text{Order } \sigma^0: & \nabla H_0 = 0 \\ \text{Order } \sigma^1: & \nabla^2 H_1 + \nabla g \cdot \underline{\nabla} H_0 = 0 \\ \quad \vdots & \quad \vdots \end{cases}$$

Comparing the zero-order equation to the mean equation (3.5a) shows that H_0 is just an approximation for the mean head $\langle H \rangle$, *to first order in* σ. Likewise, the first order term (σH_1 in the series expansion) is just an approximation for the head perturbation h, *to second order in* σ. Accordingly, we obtain:

$$\nabla^2 \langle H \rangle = 0 \ (\sigma)$$

$$\nabla^2 h + \underline{\nabla} f \cdot \underline{\nabla} \langle H \rangle = 0(\sigma^2).$$

These equations appear tractable if the high order terms on the right-hand side are neglected (requiring σ small), and the mean hydraulic gradient is constant. Note that $\langle \underline{\nabla} H \rangle$ is indeed constant to first order. This can be seen by observing that the mean head satisfies the Laplace equation to first order:

$$\nabla^2 \langle H \rangle = 0$$

which yields a linear solution for $\langle H(x) \rangle$ if the flow domain is a rectangular prism, with Dirichlet conditions on two opposite faces, and zero Neumann conditions on all other faces. As the size of the domain becomes infinite, these boundary conditions becomes equivalent to specifying a constant mean hydraulic gradient:

$$\underline{J} = - \langle \underline{\nabla} H \rangle \ . \qquad (3.6.a)$$

In practice, this kind of "boundary condition" corresponds to the case of a uniform flow field at the large scale.

The perturbative equation can now be expressed, to second order in σ, as follows:

$$\nabla^2 h - \underline{J} \cdot \underline{\nabla} f = 0 \qquad (3.6.b)$$

where the term $\underline{J}\cdot\nabla f(\underline{x})$ is a known random field. Note that equation (3.6b) is a stochastic Poisson equation governing the head perturbation $h(\underline{x})$.

We proceed to solve the stochastic Poisson equation (3.6b) in the Fourier space, by using Fourier-Stieltjes representations for both $f(x)$ and $h(\underline{x})$ as shown below:

$$f(\underline{x}) = \iiint_{-\infty}^{+\infty} e^{j\underline{k}\underline{x}} \, dZ_f(\underline{k})$$

$$h(\underline{x}) = \iiint_{-\infty}^{+\infty} e^{j\underline{k}\underline{x}} \, dZ_h(\underline{k}). \qquad (3.7)$$

Such a representation exists and is unique for any zero-mean *stationary* random field (Yaglom 1962; Loève 1963). In particular, this implies that the two-point covariance function depends only on the separation vector between the two points. By writing the representation (3.7) for $h(\underline{x})$, we therefore assume that $h(\underline{x})$ is stationary, i.e., statistically invariant under translation. The validity of this assumption will be discussed at a later stage.

The usefulness of the representation (3.7) lies in the fact that it is "orthogonal". The random dZ terms are zero-mean

complex Fourier-Stieltjes increments which have the property of being uncorrelated for distinct wavenumbers:

$$\langle dZ(\underline{k})\ dZ^{*}(\underline{k}')\rangle = 0 \quad \text{for} \quad \underline{k} \neq \underline{k}'.$$

In addition, the real and imaginary parts of $dZ(\underline{k})$ are uncorrelated and identically distributed; the real part is even in $\underline{k}$, while the imaginary part is odd:

$$dZ(-\underline{k}) = dZ^{*}(\underline{k}) \quad .$$

Finally, the variance of $|dZ|$ corresponds to the spectral content of fluctuations occurring in the wavenumber range $(\underline{k}, \underline{k}+d\underline{k})$. More precisely, it is easily seen that:

$$\langle |dZ(\underline{k})|^{2} \rangle = S(\underline{k})d\underline{k}$$

where $S(\underline{k})$ is the spectral density, i.e., the Fourier transform of the covariance function $R(\underline{\xi})$ (see equation 3.10 below).

Plugging (3.7) into (3.6) yields, by the uniqueness of the spectral representations, a simple relation between the complex Fourier increments of $h(\underline{x})$ and $f(\underline{x})$:

$$dZh\ (\underline{k}) \cong - j \frac{(J_{\ell}.k_{\ell})}{k^{2}} dZ_{f}\ (\underline{k}). \qquad (3.8)$$

Multiplying both sides by $dZ_f^*(\underline{k})$ and averaging, gives the spectral density of $h(\underline{x})$ as a function of that of $f(\underline{x})$:

$$\boxed{S_{hh}(\underline{k}) = \frac{(J_\ell k_\ell)^2}{k^4} S_{ff}(\underline{k})} \qquad (3.9)$$

where we used Einstein's implicit summation over repeated indices, and k is the radial wavenumber $\sqrt{k_1^2+k_1^2+k_3^2}$. Finally, the covariance function of $h(\underline{x})$ can be obtained by an inverse Fourier Transform of its spectrum:

$$R_{hh}(\underline{\xi}) = \iiint_{-\infty}^{+\infty} e^{j.\underline{k}.\underline{\xi}} S_{hh}(\underline{k})\, d\underline{k} \qquad (3.10)$$

The spectral densities for the *head gradient* obtain easily from the well-known relations between a stationary field and its derivative. Denoting h_i the ith-component of $\underline{\nabla}h$:

$$dZ_{h_i}(\underline{k}) = j\, k_i\, dZ_h(\underline{k}) = k_i \frac{(J_\ell k_\ell)}{k^2}\, dZ_f(\underline{k}) \qquad (3.11)$$

which gives the spectral density tensor of the head gradient vector(h_i):

$$\boxed{S_{h_i h_j}(\underline{k}) = \frac{k_i\, k_j\, (J_\ell . k_\ell)^2}{k^4} S_{ff}(\underline{k})} \qquad (3.12)$$

The tensor of covariance functions $R_{h_i h_j}(\underline{\xi})$ could be obtained by Fourier-transforming the tensor spectrum in (3.12). Alternatively, the same result can be obtained directly from $R_{hh}(\underline{\xi})$ by applying simple differentiation rules as shown below (let $\underline{\xi} = \underline{x}' - \underline{x}$):

$$R_{hh}(\underline{\xi}) = R_{hh}(\underline{x}, \underline{x}') = \langle h(\underline{x}) h(\underline{x}') \rangle$$

$$\frac{\partial R_{hh}}{\partial x_i} = \left\langle \frac{\partial h}{\partial x_i}(\underline{x}) h(\underline{x}') \right\rangle$$

$$\frac{\partial^2 R_{hh}}{\partial x_i \partial x'_j} = \left\langle \frac{\partial h}{\partial x_i}(\underline{x}) \cdot \frac{\partial h}{\partial x'_j}(\underline{x}') \right\rangle .$$

Using $\underline{x}' = \underline{x} + \underline{\xi}$ and the fact that R_{hh} depends only on $\underline{\xi}$, this leads to:

$$\frac{\partial^2 R_{hh}}{\partial \xi_i \partial \xi'_j} = - \left\langle \frac{\partial h}{\partial x_i}(\underline{x}) \cdot \frac{\partial h}{\partial x'_j}(\underline{x}+\underline{\xi}) \right\rangle .$$

Thus, the head gradient covariance tensor is simply given by:

$$\boxed{R_{h_i h_j}(\underline{\xi}) = - \frac{\partial^2 R_{hh}(\underline{\xi})}{\partial \xi_i \, \partial \xi_j}.} \qquad (3.12')$$

Finally, we follow Gelhar and Axness (1983) by using the local Darcy equation to obtain the spectral density tensor of the *flux vector*. It turns out that an additional approximation is needed in order to obtain this result. Indeed, the Darcy equation must first be expressed in terms of the log-conductivity $f(\underline{x})$:

$$\underline{Q}(\underline{x}) = - K_G \cdot e^{f(\underline{x})} \cdot \underline{\nabla} H. \qquad (3.13)$$

The exponential dependence on $f(\underline{x})$ is the source of the trouble, since the spectral representation method is only useful when the random fields appear linearly. *Gelhar and Axness* (1983) propose linearizing the exponential around $f(\underline{x}) = 0$, although we will see later that this additional approximation could be avoided. Following for now the method of Gelhar and Axness, let us compute the flux moments based on the "linearized" Darcy equation. Using the expansion $e^f = 1 + f + f^2/2 + \cdots$ gives:

$$\underline{Q}(\underline{x}) \simeq - K_G \, (-\underline{J} + \underline{\nabla} h - \underline{J} \cdot f + f \cdot \underline{\nabla} h - \underline{J} \cdot f^2/2).$$

The mean flux is obtained by taking the ensemble average:

$$\langle \underline{Q} \rangle = + K_G \left[-\langle f . \underline{\nabla} h \rangle + \underline{I}(1 + \sigma_f^2/2) \right].$$

The ensemble average term $\langle f \cdot \underline{\nabla} h \rangle$ can be worked out by using the relation between dZ_{h_i} and dZ_f given in (3.11). From the spectral representation theorem (3.7) we have:

$$\langle f(\underline{x}) \cdot \frac{\partial h(x)}{\partial x_i} \rangle = \int_{-\infty}^{+\infty} \iiint \langle dZ_f \cdot dZ^*_{h_i} \rangle$$

$$= \int_{-\infty}^{+\infty} \iiint \frac{k_i (J_j k_j)}{k^2} \langle dZ_f \, dZ_f^* \rangle$$

$$= \int_{-\infty}^{+\infty} \iiint \frac{k_i (J_j k_j)}{k^2} S_{ff}(\underline{k}) d\underline{k}$$

so that we can express the mean flux $\langle Q_i \rangle$ in terms of an effective conductivity tensor $\hat{K}_{ij}$ as follows:

$$\boxed{\begin{aligned} \bar{Q}_i &= \hat{K}_{ij} \cdot \bar{J}_j \\ \hat{K}_{ij} &= K_G \left[- \int_{-\infty}^{+\infty} \iiint \frac{k_i k_j}{k^2} S_{ff}(\underline{k}) \, d\underline{k} + (1 + \frac{\sigma_f^2}{2}) \, \delta_{ij} \right] \end{aligned}} \qquad (3.14)$$

The effective conductivity $\hat{K}_{ij}$ given in (3.14) is a *second rank symmetric tensor*, provided that the spectral density

function $S_{ff}(\underline{k})$ is even in each of the wavenumber components k_i, e.g. in the case of ellipsoidal anisotropy. However, this holds only as a first order approximation, implying that the tensorial property may not hold for large values of σ_f. Equation (3.14) was obtained by Gelhar and Axness [1983-Eq. 52], who developed close-form expressions for $\hat{K}_{ij}$ in specific cases.

In order to obtain also the *flux spectrum*, we need to expand again the e^f term as explained earlier (after Gelhar and Axness). The flux perturbation equation obtains by subtracting the mean:

$$\underline{Q}(\underline{x}) = K_G\,[1+f+f^2/2+\cdots][-\underline{\nabla}h+\underline{I}]$$
$$\langle Q(x)\rangle = K_G\,[(1+\sigma_f^2/2+\cdots)\underline{I} - \langle f.\underline{\nabla}h\rangle +\cdots]$$

which gives for the flux-perturbation:

$$\underline{q}(\underline{x}) = \underline{Q}-\langle\underline{Q}\rangle = K_G \cdot\{[(1+f+f^2/2+\cdots) - (1+\sigma_f^2+\cdots)]\,\underline{I}$$
$$- \underline{\nabla}h(1+f+f^2/2+\cdots) + (\langle f.\underline{\nabla}h\rangle+\cdots)\}$$

where the dots represent higher order terms. By neglecting perturbations of products, such as $[f^2/2 - \sigma_f^2/2]$, $[f\cdot\underline{\nabla}h\rangle]$, and all higher order perturbations as well, we obtain the "first order approximation":

$$\underline{q}(\underline{x}) \approx K_G \{\underline{J}\cdot f(\underline{x}) - \underline{y}h\}. \qquad (3.15)$$

For completeness, let us also write a third order approximation, of $\underline{q}(\underline{x})$ excluding only terms like $[f^4-\langle f^4\rangle]$, etc. This gives:

$$\underline{q}(\underline{x}) \approx K_G \cdot \{\underline{J}f - \underline{y}h + \langle \underline{y}h\cdot f\rangle - \underline{y}h\cdot f + (f^2-\langle f^2\rangle)\underline{J}/2 + \underline{J}f^3/6 + \langle \underline{y}h\cdot f^2/2\rangle - \underline{y}h\cdot f^2/2\}. \qquad (3.16)$$

Using the flux perturbation approximation (3.15) as in Gelhar and Axness leads finally to the spectral density tensor. By the representation theorem (3.7) and previous results, equation (3.15) gives:

$$\boxed{S_{q_iq_j}(\underline{k}) \approx K_G^2 \cdot J_m J_n (\delta_{im} - k_i\cdot k_m/k^2)\cdot (\delta_{jn} - k_j\cdot k_n/k^2)\cdot S_{ff}(\underline{k}).} \qquad (3.17)$$

This is a second rank symmetric tensor, being the spectral density tensor of a vector whose components are stationary random fields.

In view of the results obtained so far, it is worth noting that the flow field appears to be inherently anisotropic. Indeed, the spectral density *functions* S_{hh} and $S_{q_iq_j}$ are generally anisotropic (and so is the *tensor* $S_{q_iq_j}$), even in the

case where the input spectrum S_{ff} is isotropic.

We end this section by specializing the spectral solutions (3.9) - (3.17) for the case where the log-conductivity spectrum is of "ellipsoidal type" (as defined by Van Marcke, 1983). In this important case, the spectrum and covariance of the log-conductivity are of the form:

$$S_{ff}(\underline{k}) = S_{ff}(\lambda_i^2 k_i^2)$$

$$R_{ff}(\underline{\xi}) = R_{ff}(\lambda_i^{-2} \xi_i^2)$$

in the coordinate system coinciding with the axes of statistical anisotropy of $f(\underline{x})$. In this case, recall that the effective conductivity $\hat{K}_{ij}$ obtained from the first order analysis is a symmetric tensor.

Let us focus in particular on the case where the mean head gradient is aligned with one of the principal axes of $f(\underline{x})$. Accordingly, let the x_1 axis coincide with the mean head gradient vector $\underline{J}$ and also with the principal axis corresponding to the principal value λ_1 (λ_1 is the correlation scale along x_1, in a sense to be precised later). The spectral solutions (3.9) - (3.17) can now be written as follows:

$$\boxed{\begin{aligned} S_{hh}(\underline{k}) &= \frac{J_1^2 k_1^2}{k^4} S_{ff}(\underline{k}) \\ S_{h_i h_j}(\underline{k}) &= \frac{k_i k_j \; J_1^2 k_1^2}{k^4} \; S_{ff(\underline{k})} \\ S_{q_i q_j}(k) &= K_G^2 J_1^2 \left[\delta_{i1} - \frac{k_i k_1}{k^2}\right] \left[\delta_{j1} - \frac{k_j k_1}{k^2}\right] S_{ff}(\underline{k}) \end{aligned}} \qquad (3.18)$$

where the indices vary from 1 to m (the dimension of space). In addition, the effective conductivity tensor is now diagonal; this comes from the fact that the ellipsoidal spectrum S_{ff} is even in each of the wavenumber components, so that the integral in (3.14) vanishes for $i \neq j$. Following Gelhar and Axness (1983), equation (3.14) gives in this case:

$$\boxed{\begin{aligned} \hat{K}_{ij} &= 0 \text{ for } i \neq j \\ \hat{K}_{ii} &= K_G[-\sigma_f^2 \, g_{ii} + (1+\sigma_f^2/2)] \approx K_G \cdot \exp[\sigma_f^2(1/2 - g_{ii})] \\ g_{ii} &= \int_{-\infty}^{+\infty}\!\!\!\int\!\!\!\int \frac{k_i^2 \, S_{ff}(\underline{k})}{k^2 \sigma_f^2} \, d\underline{k} \end{aligned}} \qquad (3.19)$$

Note that the exponential formula for $\hat{K}_{ii}$ in (3.19) was proposed by Gelhar and Axness, 1983, after examination of special cases of quasi one-dimensional flow for which the effective conductivity is known exactly.

In order to obtain more concrete results, such as the covariance functions of heads and fluxes we need to introduce a specific model for the spectral density function of the log-conductivity field. There exists a wide class of ellipsoidal spectral density functions, but we will see that certain spectra lead to divergence of second order moments of the flow solution, thus violating the stationarity assumption. This is discussed in the next section below.

3.2 Discussion of Admissible Log-conductivity Spectra

In this section, we analyze certain restrictions for admissible ℓn K-spectra based on results obtained in the literature. Table 3.1 summarizes some of the input log-conductivity spectra and the corresponding covariance functions used in the literature for 2- and 3-dimensional stochastic flow problems. These particular spectra were chosen for several reasons:

(i) Some of the spectra in Table 3.1 were fitted to field data; for instance *Bakr* (1976) observed a good fit between the 1D marginal spectrum obtained by integrating the 3D Isotropic Markov Spectrum over two wavenumber components, and the 1D Spectral density of the log-conductivity measured at a borehole;

Table 3.1: Isotropic and Anisotropic Spectra for 2 and 3-Dimensional Random Log-Conductivity Fields

Spectral Model	SPECTRAL DENSITY FUNCTION S_{ff}	COVARIANCE FUNCTION R_{ff}	REFERENCES
3D Ellipsoidal Markov	$\frac{\sigma^2 \ell_1 \ell_2 \ell_3}{\pi^2 \cdot (1+u^2)^2}$	$\sigma^2 \quad e^{-s}$	Bakr et al. 1978 and Gelhar and Axness, 1983
3D Ellipsoidal Hole-Markov	$\frac{4\sigma^2 \ell_1 \ell_2 \ell_3 \cdot u^2}{3\pi^2 (1+u^2)^3}$	$\sigma^2(1-s/3) \cdot e^{-s}$	Naff 1978 and Vomvoris, 1986
3D Anisotropic Hole-Markov (non-ellips.)	$\frac{4\sigma^2 \ell_1 \ell_2 \ell_3 u_3}{\pi^2 (1+u^2)^3}$	$\sigma^2 \cdot (1-s_3^2 s) \quad e^{-s}$	Naff 1978 and Gelhar and Axness, 1983
2D Ellipsodal Markov	$\frac{\sigma^2 \ell_1 \ell_2}{\pi (1+u^2)^2}$	$\sigma^2 \cdot s \cdot K_1(s)$ (Bessel Function K_1)	Mizell et al. 1982 ($\ell_1 = \ell_2$)
2D Ellipsodal Hole-Markov	$\frac{2\sigma^2 \ell_1 \ell_2 u^2}{\pi (1+u^2)^3}$	$\sigma^2 \{\frac{s}{2} \cdot K_1(\frac{s}{2}) - \frac{s^2}{4} K_0(\frac{s}{2})\}$	Mizell et al. 1982 ($\ell_1 = \ell_2$)

Note: u_i represents the rescaled wavenumber $\lambda_i k_i$ (here without summation) and $u^2 = u_1^2 + u_2^2 + u_3^2$: similarly s_i represents the rescaled separation vector $s_i = \xi_i / \lambda_i$, and $s^2 = s_1^2 + s_2^2 + s_3^2$.

(ii) The proposed spectra are fairly simple rational functions of the wavenumber, so that closed form solutions can be obtained for at least some of the statistics of interest, such as variances and covariance functions;

(iii) Certain properties of the log-conductivity spectrum, in particular concerning the behaviour at zero wavenumber, are required in order to obtain physically realistic solutions for stochastic flow and convection-dispersion problems.

The last statement may require some explanation. Previous applications of the spectral solutions (3.18) in the literature have shown that certain quantities of interest may go to infinity (see references in Table 3.2). The divergence problem manifests itself by the appearance of a divergent integral, the integrand being typically the product of the log-conductivity spectrum by a certain transfer function which depends on the statistical quantity of interest. Such divergences are in fact a common problem in statistical physics. Two different types of divergence may be distinguished, as explained below:

(i) Low-Wavenumber divergence:

The statistical quantity of interest may diverge because of a singularity of the spectral integrand at zero wavenumber. This has been dubbed "Infrared Catastrophe". Physically, such divergence is caused by the persistence of fluctuations at increasingly *large scales* (low wavenumbers).

(ii) Large-Wavenumber divergence

On the other hand, certain statistical quantities diverge as the spectral integral is carried out to infinite wavenumbers, because the integrand does not decay rapidly enough at large wavenumbers. This type of divergence is known as "Ultraviolet Catastrophe". Physically, this means that the statistical quantity diverges because of the persistence of fluctuations at infinitely *small scales* (large wavenumbers).

Now, such divergence problems do occur in certain cases with the infinite domain spectral theory of stochastic flow and solute transport. Table 3.2 summarizes some of the results obtained in the literature, along with the relevant references. The table shows that certain constraints on the log-conductivity spectrum are needed in order to avoid the *low-wavenumber divergence* of hydraulic heads (for steady saturated flow) and of concentrations (for steady solute transport in a steady flow).

Table 3.2: Admissible Log-Conductivity Fields for Steady Flow in 1, 2, 3 Dimensions and 3D Solute Transport(*)

	CONDITION ON COVARIANCE (REFERENCES)	EXAMPLE OF ADMISSIBLE $\ell n\ K$	INTERPRETATION
1D flow ($W=1/K$)	$\int_0^\infty \xi \cdot R_{WW}(\xi)d\xi<0$ Gutjahr,Gelhar 1981	$R_{WW}(\xi) =$ $\sigma_W^2 \ (1-\frac{\lvert\xi\rvert}{\ell})e^{-\lvert\xi\rvert/\ell}$ 1D Hole-Exponential Covariance	Requires strongly negative hole-covariance with negative integral scale.
2D flow ($f = \ell n\ K$ isotropic)	$\int_0^\infty \xi \cdot R_{ff}(\xi)d\xi<0$ Mizell, Gutjahr, and Gelhar, 1982	$S_{ff}(\underline{k}) =$ $\frac{2\sigma_f^2\alpha^2}{\pi} \cdot \frac{k^2}{(k^2+\alpha^2)^3}$ 2D Markov Spectrum with Multidirectional Hole	Requires weakly negative hole-covariance, or hole-spectrum with zero integral scale.
3D flow ($f = \ell n\ K$ isotropic)	$\int_0^\infty \xi \cdot R_{ff}(\xi)d\xi<\infty$ Gelhar, Axness 1983 Gutjahr,Gelhar 1981	$S_{ff}(\underline{k}) =$ $\cdot \frac{\sigma_f^2\lambda^3}{\pi^2} + \frac{1}{(1+\lambda^2k^2)^2}$ 3D Gauss-Markov Spectrum	No Hole Spectrum is required. Any 3D isotropic field with $R_{ff}(\xi) \geq 0$ and finite integral scale is satisfactory.
3D Solute Transport ($f = \ell n\ K$ isotropic)	$S_{ff}(0) = 0$, i.e.: $\int_0^\infty \xi^2 \cdot R_{ff}(\xi)d\xi = 0$ Vomvoris 1986	$S_{ff}(\underline{k}) = \frac{4\sigma_f^2\lambda^3}{3\pi^2} \frac{\lambda^2k^2)^3}{(1+\lambda^2k^2)^3}$ 3D Gauss-Markov with Multidirectional Hole	Requires Hole-spectrum with with zero integral scale

(*) The table gives necessary and sufficient conditions for stationary solutions of the 1D, 2D, 3D flow problems with an isotropic $\ell n\ K$ field. For the 3D solute transport problem the given condition is sufficient, but may not be necessary.

In order to understand the requirements of Table 3.2, it is useful to note the relation between the integral correlation scale and the spectral density at zero wavenumber. For an m-dimensional isotropic random field (Vanmarcke, 1983) the correlation length λ_{mD} satisfies:

$$(\lambda_{mD})^m = c' \cdot \int_0^\infty \xi^{m-1} \frac{R(\xi)}{\sigma_f^2} d\xi = c'' \cdot \frac{S(0)}{\sigma_f^2}$$

Now, *Table* 3.2 shows that in order to avoid the low-wavenumber (large scale) divergence of the *head field*, a zero integral scale of ℓn K is required in the 2D case. In 3D, the requirement is much milder: it is easily shown for instance that if then any ℓn K field with finite correlation scale and positive covariance function will be satisfactory. Finally, in 1D, the requirement is much more stringent: the ℓn K correlation function must have strongly negative values and its "integral scale" must usually be negative. It has no physical meaning in this case.

Such requirements on the shape of the input spectrum could be viewed as an artefact of the *infinite domain* spectral theory. They can be explained by observing that the head field appears as the result of band-pass filtering of the ℓn K field

in the low-wavenumber range [see Figure 6.12 of Bakr, 1976]. When the flow is constrained, e.g., when the dimensionality of space is decreased from 3D to 2D or 1D, the large scale fluctuations taking place in the remaining dimensions are effectively amplified, with significant fluctuations persisting at infinitely large scales. This leads to divergent solutions in the case of steady flow in low-dimensional space (consider for instance a 3D flow channelized between two impervious walls). Such divergence can only be avoided by eliminating the large scale fluctuations of the conductivity, e.g., using "Hole-Spectra" as those shown in Table 3.1.

Table 3.2 also shows that a low-wavenumber divergence of the *solute concentration* field will occur in 3D, unless the log-conductivity field had a zero correlation scale (Vomvoris, 1986, p. 45). This shows that the problem of divergence due to large scale fluctuations may arise even if the fully 3-dimensional nature of space is taken into account (here for the solute transport problem). One possible interpretation is that the large scale divergence of solute concentrations results from the inadequate *assumption* that there exist *steady* solutions which are *spatially stationary*. Some insight on this question could be

gained by examining the fully non-steady flow/transport problem in three dimensions. This will not be attempted here.

We conclude that the *low-wavenumber or large-scale divergence* of stochastic quantities may result from the inadequacy of the assumption that there exist stationary and ergodic solutions to the flow and transport problems in the *steady state*. The results of the first order spectral perturbations theory (Table 3.2) clearly show that this is not always the case, depending on the spatial structure of the porous medium. In general, the ℓn K field must be taken to be "almost periodic" at some large but finite scale in order to obtain useful results. Furthermore, it turns out that the ℓn K field must be taken more strongly "periodic", with smaller "wavelength", as the dimensionality of space decreases. In our view, this indicates that the conditions for the existence of steady stationary solutions become more and more restrictive as the degree of freedom of flow decreases in physical space.

Finally, let us briefly mention the appearance of *large wavenumber divergences* of the steady solute concentration field, as this may be relevant to the steady flow problem itself. One example of small scale divergence is the divergence of the concentration variance as the local dispersion coefficient (length scale) goes to zero. A second case of small scale

divergence or near-divergence was found by Vomvoris, 1986: the correlation scale of the concentration along the mean flow direction appeared to be very sensitive to the rate of decay of the ℓn K spectrum at large wavenumbers. In order to obtain physically reasonable results, the ℓn K spectrum must decay *exponentially* at large wavenumbers. In our view, this suggests that the ℓn K spectrum should be truncated at some large wavenumber $k \simeq \alpha^{-1}$, where α represents the length scale of the mechanical dispersion process taking place at small scales. In other words, α represents the small scale dispersivity, or sub-grid dispersivity with respect to the measurement grid. Accordingly, we argue that the spectral density models of Table 3.1 should be truncated at $k \lesssim \alpha^{-1}$ for use in the solute transport equations. The proposed approach assumes that the effect of small scales heterogeneities is correctly modeled by the local dispersivity term, which should be measured independently at the laboratory scale.

Based on the previous analysis, both the Markov and Hole-Markov spectra appear to be admissible for stationary first order flow solutions in an isotropic three dimensional medium. Presumably, this also holds in the anisotropic case. Thus, we develop in the next sections a number of closed form results for these spectra with various degrees of anisotropy in three dimensions. Recall however that the proposed spectra may not be

meaningful in the case of stochastic solute transport, unless they be truncated at some low and high wavenumbers. The Hole model realizes a kind of smooth truncation at low wavenumbers, but seems somewhat arbitrary. This problem suggested the idea of a finite domain approach based on band-pass self-similar spectra, eventually leading to a systematic analysis of finite size effects in a later part of this work (Section 4.4 of Chapter 4).

3.3 Head and Flux Moments for the 3D Isotropic Markov Spectrum

In the isotropic case, simple closed form results can be obtained for the head variance, head correlation function, effective conductivity, and the variance of the head gradient and flux vectors. We present all these results below. The quantities σ_h^2, $R_{hh}(\xi)$, $\hat{K}_{ii}$, were obtained by Bakr et al., 1978. The head gradient variance $\sigma_{h_i}^2$ and the flux variance $\sigma_{q_i}^2$ are easily obtained by spectral integration (see Appendices 3.A and 3.B). Finally, the flux covariance functions $R_{q_i q_j}(\xi)$ can be obtained either by numerical integration or analytically (Wendy Graham, personal communication, and Appendix 3.C). All the relevant quantities were computed by using the 3D Gauss-Markov spectrum of Table 3.1 with $\lambda_1=\lambda_2=\lambda_3$. Thus, the log-conductivity spectrum is:

$$S_{ff}(\underline{k}) = \frac{\sigma_f^2 \lambda^3}{\pi^2} \cdot \frac{1}{(1+\lambda^2 k^2)^2} \qquad (3.20)$$

The head covariance obtains from:

$$R_{hh}(\underline{k}) = \iiint e^{j\underline{k}\underline{\xi}} S_{hh}(\underline{k}) \, d\underline{k}$$

and the flux covariance from:

$$R_{q_i q_j}(\underline{\xi}) = \iiint e^{j\underline{k}\underline{\xi}} S_{hh}(\underline{k}) \, d\underline{k}$$

where the spectra appearing in the integrands were given in equation (3.18). The effective conductivity $\hat{K}_{ii}$ is similarly evaluated from (3.19). The results are the following:

(i) Hydraulic head (from Bakr et al. 1978):

$$\sigma_h = \frac{1}{\sqrt{3}} \sigma_f \lambda J_1 \qquad (3.21.a)$$

$$\begin{aligned} R_{hh}(\xi) = (\tfrac{3}{2})\sigma_h^2 \cdot \{ & (\cos^2\chi - 1).[e^{-\xi/\lambda} + 2(e^{-\xi/\lambda} - 1) \cdot (\xi/\lambda)^{-1}] \\ & + (3\cos^2\chi - 1) \cdot [\, (1 - e^{-\xi/\lambda}) . (\xi/2\lambda)^{-3} \\ & - e^{-\xi/\lambda}(1 + 2(\xi/2\lambda)^{-1} + 2(\xi/2\lambda)^{-2})] \,\} \end{aligned} \qquad (3.21.b)$$

where χ is the angle between the separation vector $\underline{\xi}$ and the mean flow direction. The corresponding

correlation function is plotted in Figure 3.1 for two angles: $\chi = 0$ (along the mean flow) and $\chi = \pi/2$ (across the mean flow). Note $\cos\chi = \xi_1/\xi$ $\xi = \sqrt{\xi_1^2+\xi_2^2+\xi_3^2}$, so that $\chi = 0$ corresponds to $\underline{\xi} = (\xi_1, 0, 0)$ and $\chi = \pi/2$ to $\underline{\xi} = (0, \xi_2, \xi_3)$.

(ii) Head gradient vector (Appendices 3.A and 3.C):

$$\sigma_{h_1} = \frac{1}{\sqrt{5}} \sigma_f \, J_1$$

$$\sigma_{h_2} = \sigma_{h_3} = \frac{1}{\sqrt{15}} \sigma_f J_1 \qquad (3.22)$$

The head gradient covariance functions $R_{h_i h_j}(\underline{\xi})$ can be obtained as indicated in Appendix 3.C.

(iii) Mean flux and effective conductivity (from Gelhar and Axness, 1983):

$$\hat{K}_{ii} = K_G \cdot \exp(\sigma_f^2/6) \quad \forall i = 1,2,3 \qquad (3.23)$$

$$\langle Q_1 \rangle = \hat{K}_{11} \cdot J_1, \qquad \langle Q_2 \rangle = \langle Q_3 \rangle = 0.$$

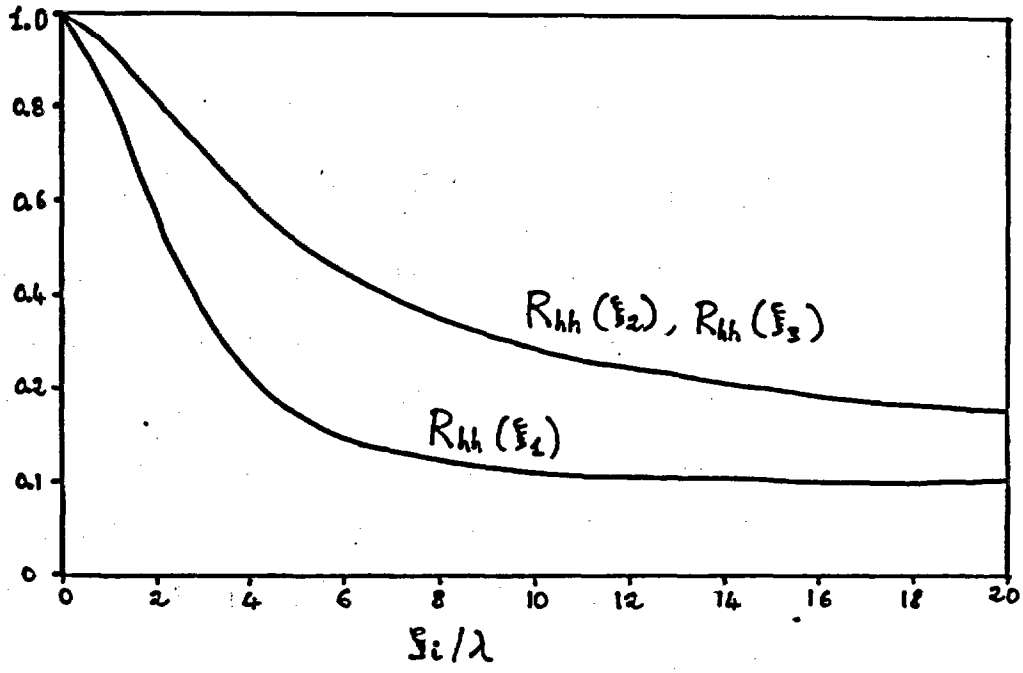


Figure 3.1 Head correlation function along the coordinate axes for the 3D Isotropic Markov spectrum of log-conductivity

(iv) Flux vector variances and covariance functions (cf. Appendices 3.B and 3.C)

$$\sigma_{q_1} = \sqrt{8/15}\ K_G\ \sigma_f\ J_1 \qquad (3.24)$$

$$\sigma_{q_1} = \sigma_{q_2} = \sqrt{1/15}\ K_G\ \sigma_f\ J_1$$

$$R_{q_i q_j}(\underline{0}) = 0 \quad \text{for } i \neq j$$

The covariance tensor $R_{q_i q_j}(\underline{\xi})$ is defined by:

$$R_{q_i q_j}(\underline{\xi}) = \langle (Q_i(\underline{x}) - \bar{Q}_i) \cdot (Q_j(\underline{x}+\underline{\xi}) - \bar{Q}_j) \rangle$$

The components of this tensor are plotted along the three coordinates axes ($\underline{\xi}$ parallel to $\underline{x}_i$) in Figures (3.2) and (3.3). These plots were obtained from analytical integration of the flux spectrum of equation (3.18). The integrations were carried out by Wendy Graham (personal communication) for all but the covariance $R_{q_2 q_2}(\xi_3)$ given in Appendix 3.C.

The results in equations (3.21)-(3.24) and Figures (3.1)-(3.3) can be interpreted as follows. In terms of standard deviations, the equation for σ_H shows that the amplitude of head fluctuations is proportional to σ_f and to the mean head drop over one correlation scale (λJ_1). The amplitude of the flux vector fluctations is proportional to σ_f and to $K_G\ J_1$ (a subsequent analysis will show that this term should be replaced by the mean flux $\langle Q_1 \rangle$ -- see section 4.3). Similarly, the amplitude of fluctuations of the head gradient is

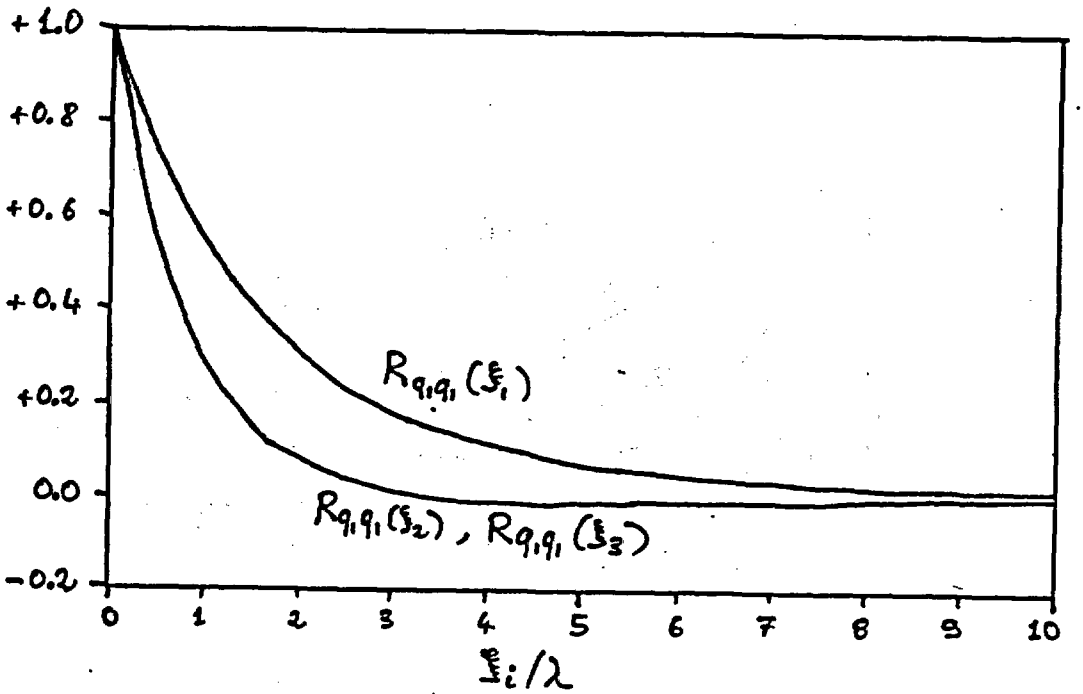


Figure 3.2 Longitudinal flux correlation function along the coordinate axes for the 3D Isotropic Markov spectrum of log-conductivity

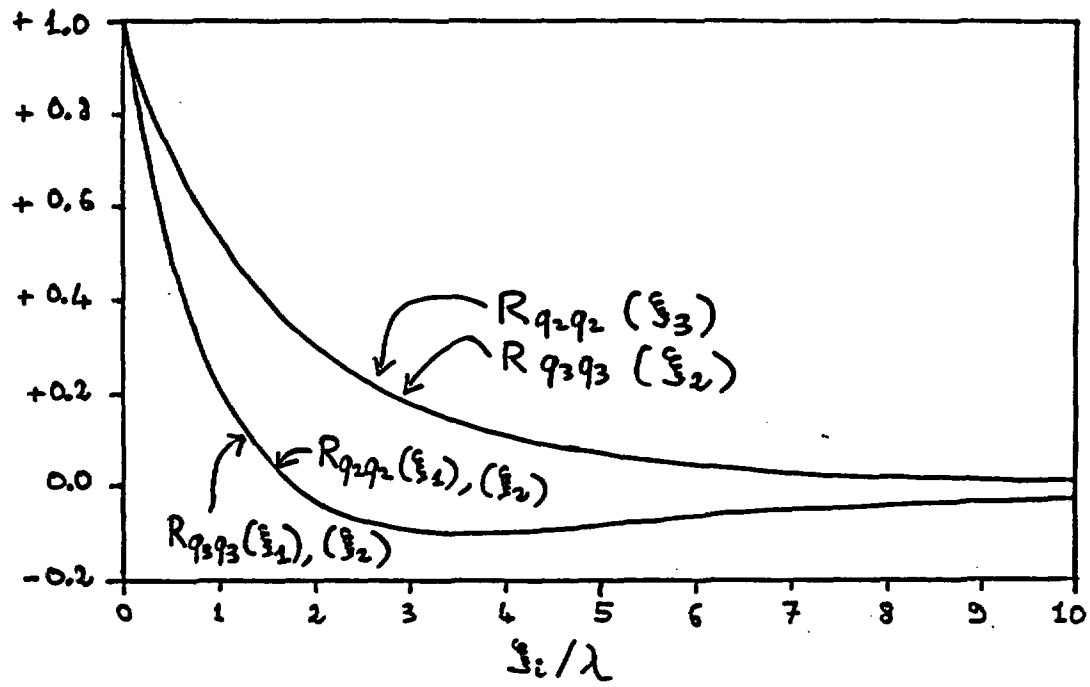


Figure 3.3 Transverse flux correlation functions along the coordinate axes for the 3D Isotropic Markov spectrum of log-conductivity

proportional to σ_f and to the mean head gradient J_1.

Furthermore, it appears that the *longitudinal* flux component has larger fluctuations than the *transverse* flux components, by a factor $\sqrt{8}$. The contrast is milder for the head gradient components (factor $\sqrt{3}$). After replacing the term $K_G J_1$ by the mean flux $\langle Q_1 \rangle$ in (3.24), we obtain a coefficient of variation on the order of $\sigma_f/\sqrt{2}$ for the longitudinal flux component. The coefficient of variation for the longitudinal component of the head gradient is significantly smaller, about $\sigma_f/\sqrt{5}$. This suggests that the range of validity of the first order solutions may not be the same for different quantities such as head gradient and flux. Taking for instance $\sigma_f \simeq 1.5$ yields a coefficient of variation of 60% for $\partial h/\partial x_1$, and over 100% for q_1. It seems wise to expect some degree of inaccuracy of the perturbative solutions for such a large coefficient of variation as 100%. This example indicates that the range of validity of the spectral solutions for the flux vector could be limited to cases of moderate variability ($\sigma_f < 1$-1.5).

In terms of correlation functions, it can be seen from Figure (3.1) that the hydraulic head is correlated over longer distances than the log-conductivity. Furthermore, the head correlation is stronger in the direction transverse to flow.

Defining the e-correlation length as the distance at which correlation drops below e^{-1}, we obtain a head correlation length equal to 7.5λ transversally, compared to 3λ in the direction of the mean flow. This clearly shows that the head field is *not* isotropic.

Similarly, the flux components are not isotropic. The flux correlation lengths are on the order of the conductivity scale λ, or nearly twice as much for certain flux components and orientations of the separation vector (see Figures (3.2) and (3.3)). Overall, it appears that the flux correlations are consistently smaller than the head correlations. This can be explained by the fact that the flux is more directly related to the conductivity fluctuations (through the local Darcy equation). Another remarkable feature is the fact that different flux components are totally uncorrelated at zero separation distance. Finally, note again that the *anisotropic* flux covariance tensor satisfies a number of symmetry relations and other identities (mass conservation), which will be analyzed in a later part of this work (Section 4.2).

We end this section by focusing on a peculiar feature of the flux spectral solution which does not seem to have been observed in the literature. According to equations (3.23) and

(3.24), it appears that the ratios $\sigma_{q_i}/\langle Q_1\rangle$ have a maximum at $\sigma_f = \sqrt{3}$, and they converge to zero as σ_f goes to infinity. Indeed, from (3.23) and (3.24) we have:

$$\frac{\sigma_{q_1}}{\langle Q_1\rangle} = \sqrt{8/15}\ \sigma_f\ e^{-\sigma_f^2/6}$$

$$\frac{\sigma_{q_2}}{\langle Q_1\rangle} = \frac{\sigma_{q_3}}{\langle Q_1\rangle} = \sqrt{1/15}\ \sigma_f\ e^{-\sigma_f^2/6} \qquad (3.25)$$

Surprisingly, these quantities have a maximum at $\sigma_f = \sqrt{3}$:

$$\left.\frac{\sigma_{q_1}}{\langle Q_1\rangle}\right|_{max} = 0.557$$

$$\left.\frac{\sigma_{q_2}}{\langle Q_1\rangle}\right|_{max} = \left.\frac{\sigma_{q_3}}{\langle Q_1\rangle}\right|_{max} = 0.271.$$

Unfortunately, there seems to be no obvious physical reason for such a behavior. Rather, one would expect that the coefficient of variation $\sigma_{q_1}/\langle Q_1\rangle$ be a monotonously increasing function of σ_f. We will show in Chapter 4 (section 4.3) that a more physical behavior obtains by an alternative perturbation analysis which avoids linearization approximations of the type $e^f \simeq 1 + f + \cdots$. The new spectral solution obtains simply by replacing the term $K_G J_1$ by $\langle Q_1\rangle$ in the flux variance (3.24).

Therefore, the exponential term $e^{-\sigma^2/6}$ disappears from (3.25), and the coefficients of variations of the flux components appear now to increase linearly with σ_f, which seems to be a more realistic behavior. Accordingly, we propose that equation (3.24) be modified as follows:

$$\boxed{\begin{aligned} \sigma_{q_1} &= \sqrt{8/15}\ \sigma_f\ \langle Q_1 \rangle \\ \sigma_{q_2} &= \sigma_{q_3} = \sqrt{1/15}\ \sigma_f\ \langle Q_1 \rangle\ . \end{aligned}} \qquad (3.26)$$

This modification of the spectral flow solutions of Gelhar and Axness (1983) has also implications on the solute transport problem, as explained later (Section 4.3 of Chapter 4).

3.4 Head and Flux Moments for the 3D Anisotropic Markov Spectrum

The first order spectral solutions (3.18) are now applied to the case of the ellipsoidal Markov log-conductivity spectrum (Table 3.1) with $\ell_1=\ell_2=\ell$ and $\ell_3 \neq \ell$. We assume that the mean head gradient is parallel to the principal direction of anisotropy (x_1). The vertical/horizontal anisotropy ratio:

$$\epsilon = \ell_3/\ell$$

will usually be taken to be less than one, as this is a case of practical interest for most horizontally layered porous media. In addition, certain results become simpler when $\epsilon \ll 1$ holds,

which is the case of perfectly stratified media. However, this assumption is not needed in the calculations that follow.

The log-conductivity spectrum for the case at hand is:

$$S_{ff}(\underline{k}) = \frac{\sigma_f^2 \ell^2 \ell_3}{\pi^2} \cdot \frac{1}{[1+\ell^2(k_1^2+k_2^2) + \ell_3^2 k_3^2]^2}$$

Plugging this into (3.10), we obtain the head covariance function in the form:

$$R_{hh}(\underline{k}) = \left[\frac{\pi}{8}\sigma_f^2 \cdot J_1^2 \, \ell\ell_3\right] \cdot I(\underline{\xi}) \qquad (3.27.a)$$

where $I(\underline{\xi})$ is the triple-integral:

$$I(\underline{\xi}) = \left[\frac{2}{\pi\ell}\right]^3 \cdot \int\!\!\int\!\!\int_{-\infty}^{+\infty} \frac{k_1^2 \cos(\underline{k}\underline{\xi})}{k^4[\ell^{-2}+(k_1^2+k_2^2)+\epsilon^2 k_3^2]^2} \, d\underline{k}$$

We show in Appendix 3.D that that this can be reduced to the double-integral:

$$I(\underline{\xi}) = \frac{2}{\Pi^2} \cdot \int_0^{2\Pi} \cos^2\theta \, d\theta \int_0^{\Pi} F(\theta,\varphi) \, d\varphi \qquad (3.27.b)$$

where

$$F(\theta,\varphi) = \sin^3\varphi \cdot \left[\frac{1+C(\theta,\varphi)}{B(\varphi)}\right] \cdot e^{-C(\theta,\varphi)}$$

$$C(\theta,\varphi) = \frac{1}{\ell} \cdot \left|\frac{A(\theta,\varphi)}{B(\varphi)}\right|$$

$$A(\theta,\varphi) = (\xi_1\cos\theta+\xi_2\sin\theta)\sin\varphi+\xi_3\cos\varphi$$

$$B(\varphi) = \sqrt{\sin^2\varphi+\epsilon^2\cdot\cos^2\varphi}$$

The double-integral above was computed by careful numerical integration, using Romberg interpolation for the most difficult inner integral (IMSL routine DCADRE), and Simpson integration for the outer integral. The calculations were carried out in double precision (64-bit words). In addition, the results were checked by using a quasi-analytical expression for $R_{hh}(0,0,\xi_3)$ in the case $\epsilon \ll 1$ (see Appendix 3.E). The comparison indicates that the numerical integration procedure was sound.

Figure (3.4) shows the resulting head covariance functions plotted for a separation vector $\underline{\xi}$ parallel to x_1, x_2 and x_3 respectively, for different values of the anisotropy ratio. Only the case $\epsilon \leq 1$ is shown, as this is the most interesting case in practice (horizontal layering).

The most remarkable feature from these plots is that

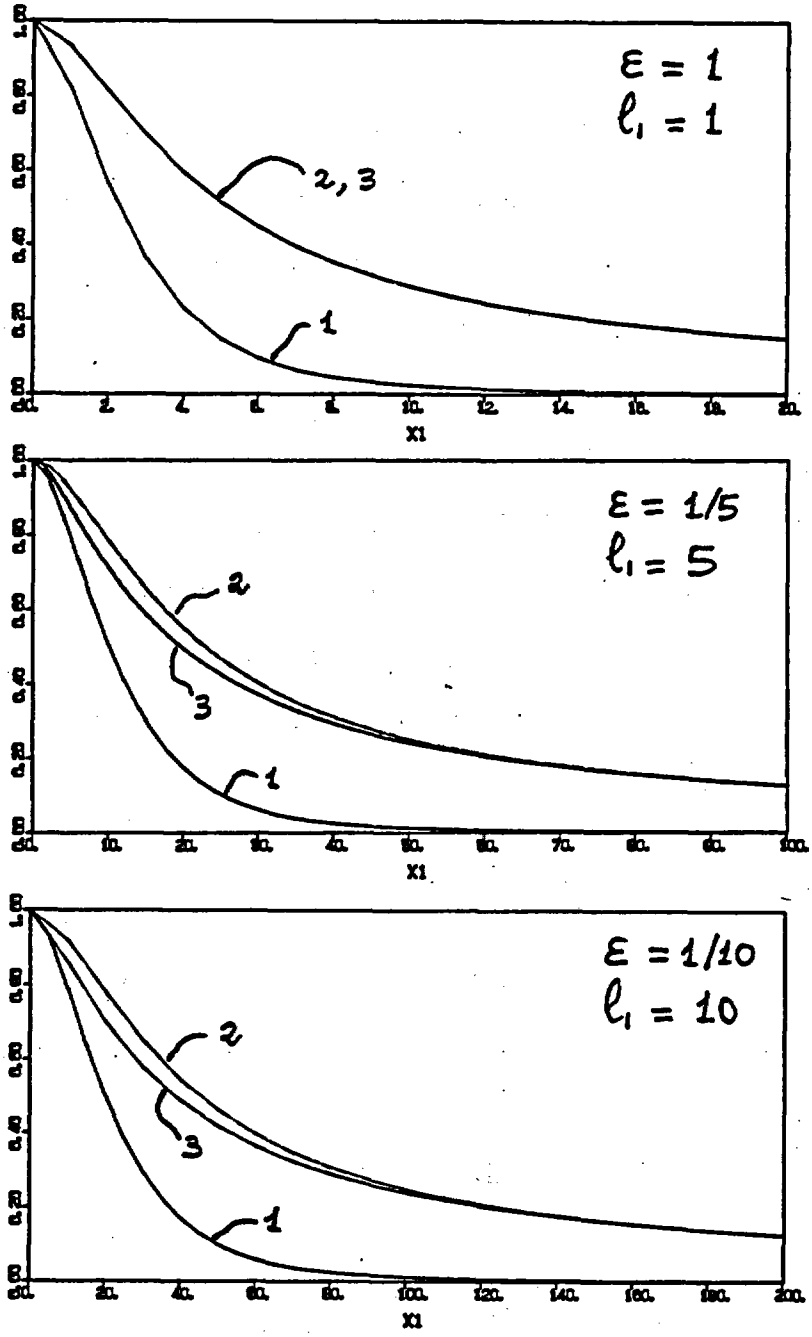


Figure 3.4 Head correlation function $R_{hh}(\underline{\xi})$ along the three principal directions for the 3D ellipsoidal Markov spectrum of log-conductivity with different values of the anisotropy ratio ($\epsilon = \ell_3/\ell_1$) as ℓ_1 increases.

the head correlation function does not change much when the vertical scale ℓ_3 goes to zero as the horizontal scale ℓ remains fixed. Indeed, the vertical correlation scale of heads only decreases slightly as ℓ_3 decreases from $\ell_3=\ell$ (isotropic) to $\ell_3\simeq 0$ (perfectly stratified) while ℓ is fixed. Physically, this means that the head process in a layered medium has a vertical scale of fluctuation much larger than the "layers" thickness, and does not decrease appreciably as the layers thickness decreases. Thus, as $\epsilon \to 0$, the ratio of the vertical head scale versus the vertical log-conductivity scale tends to infinity.

Furthermore, the ensemble head variance may be obtained analytically by evaluating the integral $I(\xi)$ at $\xi = 0$. This was computed by *Naff and Vecchia* (1986) for the case $\epsilon \leq 1$:

$$\sigma_h^2 = \frac{\pi}{8} \sigma_f^2 J_i^2 \ell_G^2 \cdot G(\epsilon) \qquad (3.28)$$

where:

$$G(\epsilon) = \frac{1}{\gamma} \{(1 - \frac{\epsilon^2}{\gamma^2}) + \frac{2}{\pi} \frac{\epsilon}{\gamma}\}$$
$$+ \frac{1}{\epsilon^2}(\frac{1}{\gamma^3} - \frac{3}{\gamma} + 2\gamma)\cdot(1 - \frac{2}{\pi} \tan^{-1} (\gamma/\epsilon))$$

and:

$$\tau = \sqrt{1-\epsilon^2}$$

$$\epsilon = \ell_3/\ell \leq 1$$

$$\ell_G = \sqrt{\ell \ell_3}.$$

It is easily seen that $G(\epsilon) \to 1$ as $\epsilon \to 0$. This gives immediately the asymptotic result for perfectly stratified media as follows:

$$\boxed{\sigma_h \approx \sqrt{\frac{\pi}{8}}\,\sigma_f\, J_1\, \ell_G \quad \text{for } \epsilon \ll 1.} \qquad (3.29)$$

Note that $\ell_G = \sqrt{\ell \ell_3}$ is the geometric mean of the horizontal and vertical correlation scales. This shows that the head standard deviation tends to a finite constant (neither zero nor infinity) if the anisotropy ratio decreases while the geometric mean correlation scale ℓ_G remains constant.

Finally, let us analyze the statistics of the *flux vector*. In the asymptotic case of perfect stratification ($\epsilon \ll 1$), a few close-form results can be obtained. Recall for instance that equation (3.19) gives the general form of the effective conductivity in the anisotropic case (after Gelhar and Axness, 1983). For $\epsilon \ll 1$, in particular, equation (3.19) yields the geometric mean and the harmonic mean, respectively, for the effective conductivity components parallel and transverse to stratification:

$$\hat{K}_{11} \simeq K_G$$
$$\hat{K}_{22} \simeq K_G \qquad (3.30)$$
$$\hat{K}_{33} \simeq K_H = K_G\, e^{-\sigma_f^2/2}.$$

We show in Appendix (3.F) that closed form expressions can be obtained as well for the variance of the flux components ($\sigma^2_{q_i}$) in the limit of perfect stratification $\epsilon \to 0$:

$$\boxed{\begin{aligned} \sigma_{q_1} &\simeq \sqrt{1 - \frac{13}{32}\pi\,\epsilon}\ \cdot\ \sigma_f\, K_G\, J_1 \\ \sigma_{q_2} &\simeq \sqrt{\frac{\pi}{32}\epsilon}\ \cdot\ \sigma_f\, K_G\, J_1 \\ \sigma_{q_3} &\simeq \sqrt{\frac{\pi}{8}\epsilon}\ \cdot\ \sigma_f\, K_G\, J_1 \end{aligned}} \qquad (3.31)$$

These approximate relations were obtained for the 3D anisotropic Markov spectrum with $\ell_1 = \ell_2 = \ell$ and assuming $\epsilon = \ell_3/\ell$ small (they are thought to be adequate for $\epsilon \lesssim 1/5$ or so).

It appears from equations (3.31) that both transverse flux variances vanish in the limit of perfect stratification, while the longitudinal flux variance tends to a constant value about twice larger than would be obtained in the isotropic case (compare equations 3.31 and 3.24). Accordingly, the transverse flux components Q_i ($i = 2,3$) appear to vanish identically when

$\epsilon = 0$, resulting in a one-dimensional flow. However, the mass conservation equation shows that Q_1 must be constant in one-dimensional flow, which seems contrary to the limit result $\sigma_{q_1} \neq 0$, $\sigma_{q_2} = \sigma_{q_3} = 0$. This difficulty can be resolved by considering the fact that the transverse flux components have smaller scales of fluctuations than the longitudinal component.

The correlation length scales of the flux components can be evaluated qualitatively by considering the mass conservation equation in relation to equation (3.31). Using the fact that the mean flux vector is constant, mass conservation simply requires that the flux perturbation be "divergence free", that is:

$$\frac{\partial q_1}{\partial x_1} + \frac{\partial q_2}{\partial x_2} + \frac{\partial q_3}{\partial x_3} = 0.$$

A standard "scale analysis" of this equation leads to:

$$\frac{\sigma_{q_1}}{\Lambda_{11}} u_1(\underline{x}) + \frac{\sigma_{q_2}}{\Lambda_{22}} u_2(\underline{x}) + \frac{\sigma_{q_3}}{\Lambda_{33}} u_3(\underline{x}) = 0$$

where Λ_{ii} is the correlation length of q_i along the x_i axis. and the $u_i(\underline{x})$ are normalized zero-mean random fields having variances on the order of unity. This formulation suggests that the constant coefficients in the above equation must be roughly

of equal magnitude:

$$\frac{\sigma_{q_1}}{\Lambda_{11}} \simeq \frac{\sigma_{q_2}}{\Lambda_{22}} \simeq \frac{\sigma_{q_3}}{\Lambda_{33}}.$$

Using the σ_{q_i}'s given in (3.31), and assuming that the longitudinal flux has a correlation length on the order of ℓ (the conductivity scale in the horizontal) leads to rough estimates for the flux correlation scales as follows:

$$\Lambda_{11} \simeq \ell$$

$$\Lambda_{22} \simeq \sqrt{\frac{\pi}{32}}\, \epsilon\, \ell \simeq 0.3\, \sqrt{\ell \ell_3} \qquad (3.32)$$

$$\Lambda_{33} \simeq \sqrt{\frac{4\pi}{32}}\, \epsilon\, \ell \sim 0.6\, \sqrt{\ell \ell_3}.$$

Thus, although the transverse flux components have very small variances, their fluctuation scale is also very small. As a consequence, the terms $\partial q_i/\partial x_i$ (i=2,3) appearing in the mass conservation equation may not be negligible. This explains why the "perfectly stratified" flow might appear nearly one-dimensional in the large, while still retaining three-dimensional features at the local scale.

The actual spatial structure of the flux vector field is probably more complex than the above analysis would suggest. However, it is sufficient here to observe that the flux vector has sharply contrasting scales of fluctuations in different directions and components, while the hydraulic head field is only slightly anisotropic (as can be seen from Figure 3.4). The practical implication of these findings for groundwater flow in stratified subsurface formations is now examined.

3.5 Discussion of the Anisotropic Case (Stratified Flow Systems)

Some of the new results developed above have direct implications for field problems, as most subsurface formations exhibit some degree of horizontal or near-horizontal layering (see Figure 3.5). The spatial structure of the flow field for statistically anisotropic conductivities was not fully understood, it seems, although stochastic solutions were available from the work of Gelhar and Axness (1983) and others. In order to illustrate our findings, we will consider the case of an aquifer with significant vertical-to-horizontal anisotropy, say $\ell_3/\ell_1 = 1/5$, and isotropy in the horizontal plane of stratification. The mean flow is assumed, as before, to be parallel to the plane of stratification. The log-conductivity spectrum is assumed to be the 3D anisotropic Markov model. For illustration, we will use typical length scale values as follows:

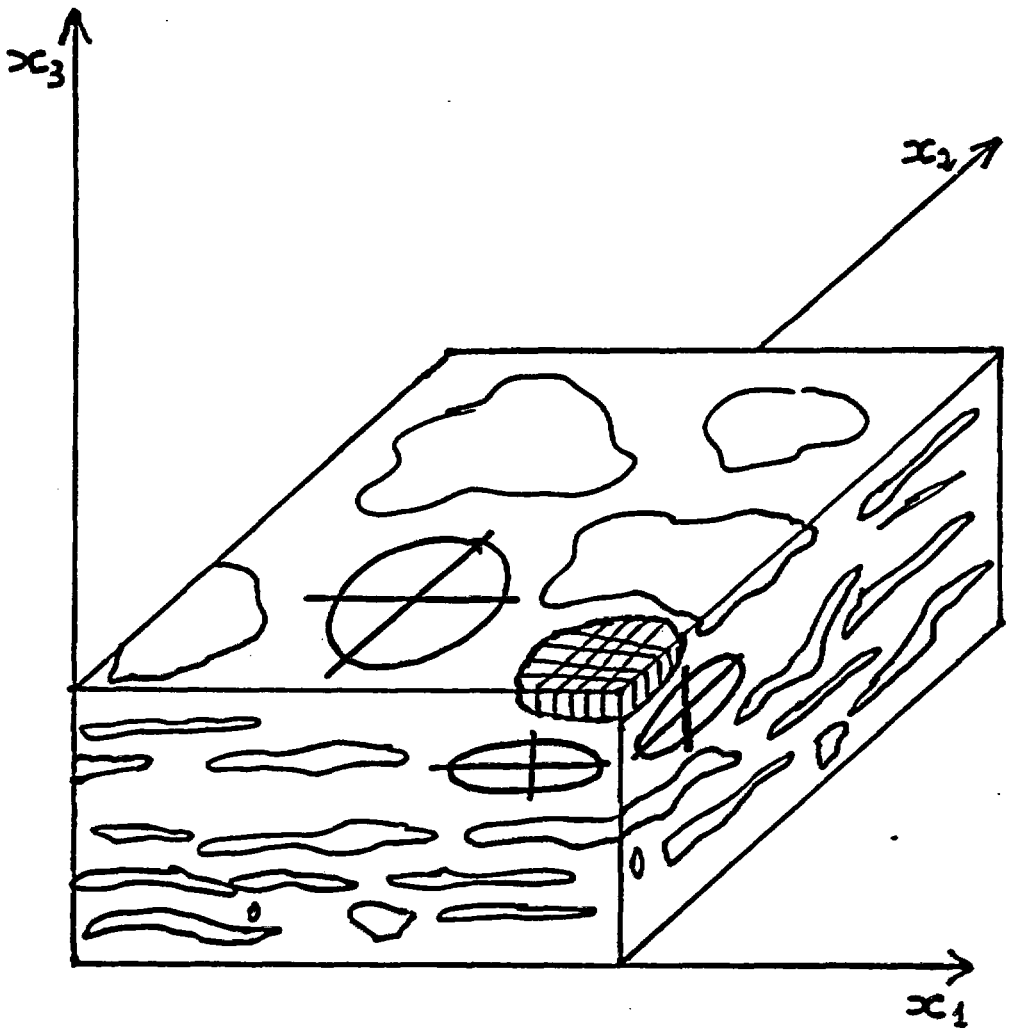


Figure 3.5: Sketch of a statistically layered porous medium. The ellipses represent contours of constant correlation length (anisotropy ellipses in different planes, or anisotropy ellipsoid in 3D space) for the log-conductivity field.

$\ell_1 = \ell_2 \simeq 1m$, $\ell_3 \simeq 0.20$ m.

We now discuss some of the most salient features of the flow field, based on the statistical analysis of spectral solutions developed in this section. Perhaps the most important point to be made here concerns the *correlation structure* of the flow field, in terms of hydraulic head and flux: see figures (3.5), (3.6) and (3.7). The hydraulic head exhibits near-perfect isotropy in the cross-flow plane (perpendicular to stratification). For the example above, the correlation scale of the head perpendicular to strata would be about 7 m, that is about 40 times larger than the conductivity scale in the same direction ($\ell_3 = 0.20m$). The correlation scale of head in the horizontal direction across flow is on the same order, about 7.5 m. Finally, the correlation scale of head along the mean flow direction is half smaller, about 3m. These numbers are very close to those obtained for the isotropic case $\ell_1 = \ell_2 = \ell_3 = 1m$. Thus, it appears that *the spatial structure of the hydraulic head is not senstitive to the anisotropy ratio.* Furthermore, for the anisotropic case at hand, the head will be very strongly correlated over distances on the order of one or a few meters. In fact, the hydraulic head should appear nearly constant vertically over a few layer thicknesses.

In contrast, we have seen that the vertical flux component q_3 has a scale of fluctuation presumably on the order of 0.20m along x_3 perpendicular to stratification. This is just the same as the conductivity correlation scale ℓ_3, or "layer thickness". Figure (3.6) shows that the transverse flux component q_2 also has a small correlation scale (0.10m) along x_2, parallel to stratification. On the other hand, the longitudinal flux q_1 presumably has a much smaller scale of fluctuation transverse to the mean flow (0.20m) than along the mean flow (1m). Thus, *the flux vector field appears strongly anisotropic* as illustrated on the bottom parts of Figures (3.6) and (3.7). This is in contrast with the near isotropic character of the head field (except for a ratio 1:2 in the horizontal plane).

Consider now a very shallow aquifer ($L_3 \lesssim 3$m) and a deep aquifer ($L_3 \gtrsim 100$m), with correlation scales $\ell_1 = \ell_2 = 1$m and $\ell_3 = 0.20$m as before. Because the vertical head correlation scale is about 7m, the head will appear nearly constant vertically in the shallow aquifer:

$$Q_3 = -K \frac{\partial H}{\partial x_3} \simeq 0. \qquad (3.33)$$

Upon vertical averaging, head field in the shallow aquifer system behaves nearly two-dimensionally, (or one-dimensionally) as aquifer thickness decreases. However, the intermediate case of

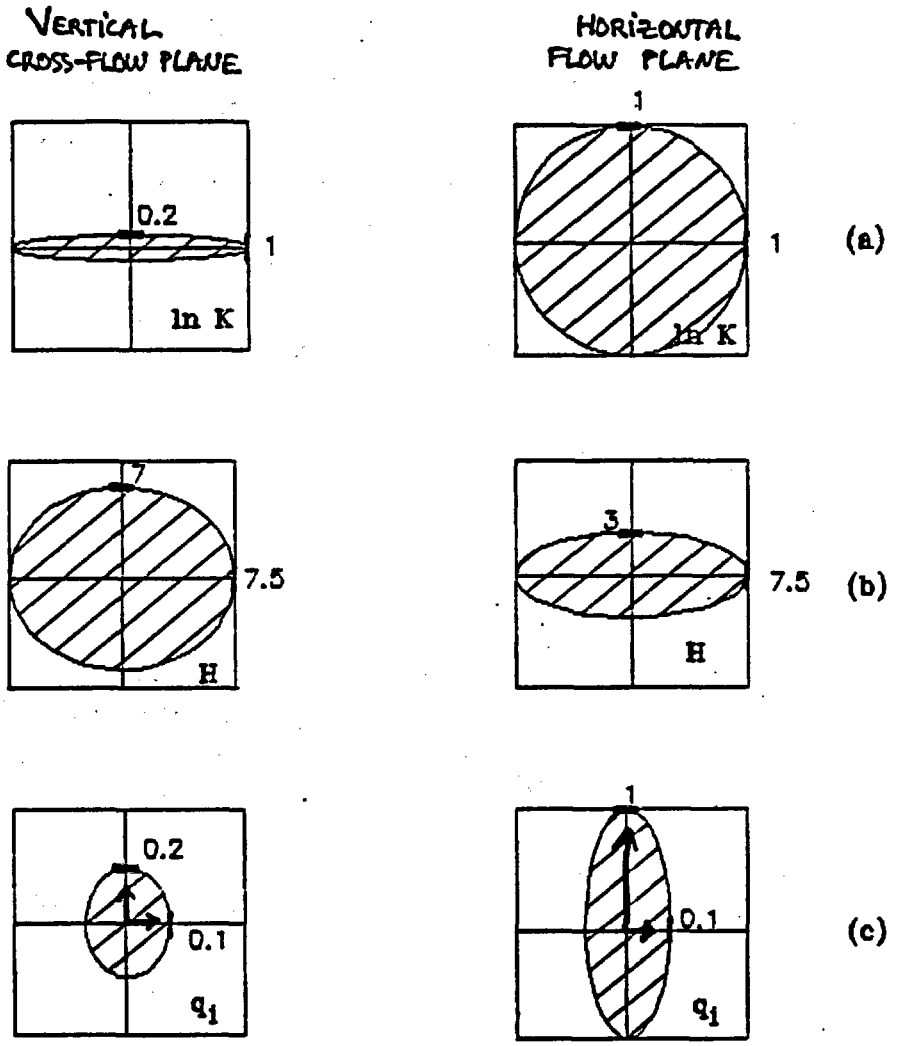


Figure 3.6: Anisotropy ellipses for (a) the log-conductivity field, (b) the hydraulic head, and (c) the flux vector in a statistically layered aquifer.

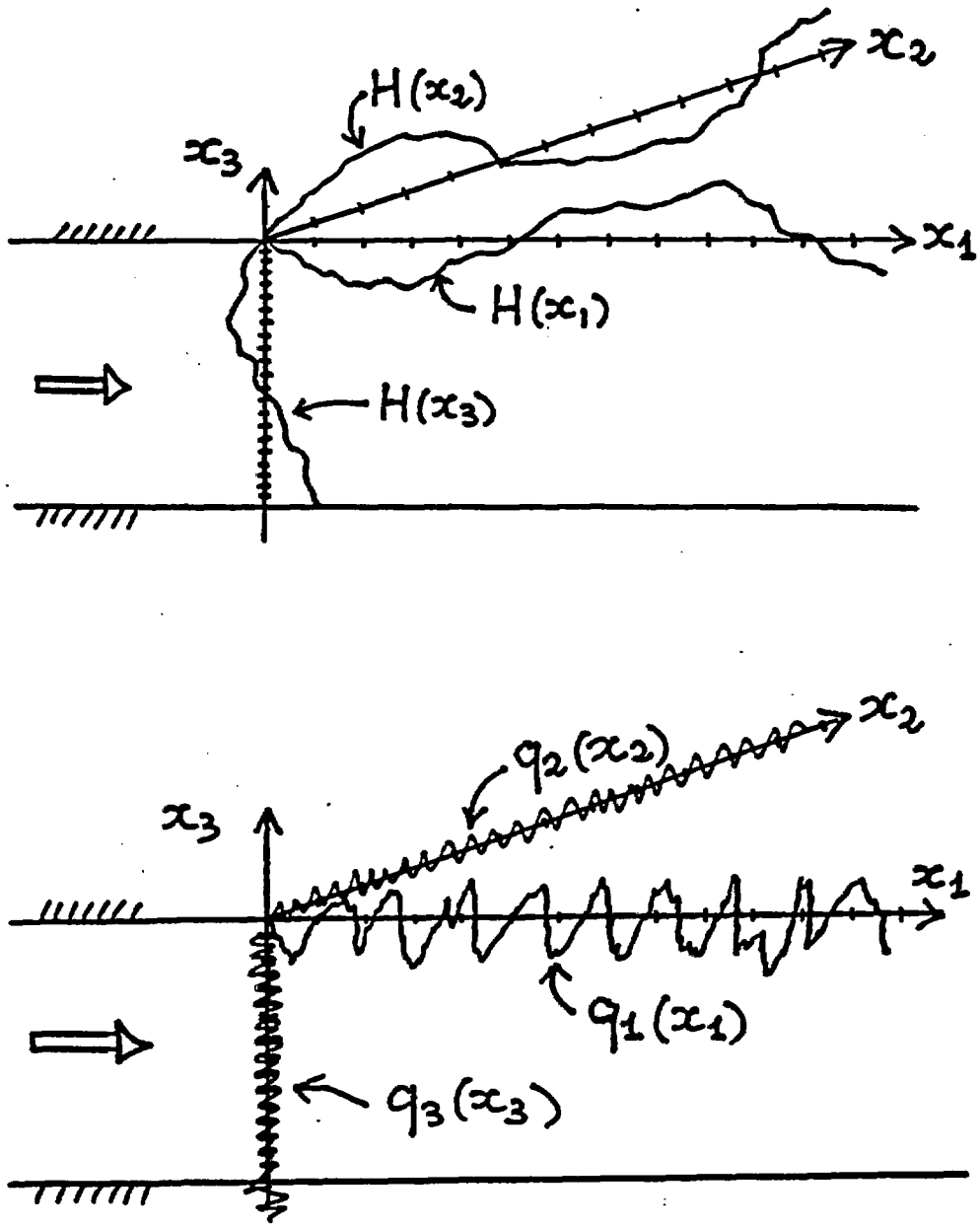


Figure 3.7: Schematic representation of the fluctuation scales of the hydraulic head (top) and of the flux vector (bottom) in a stratified aquifer

moderately shallow aquifers is more complex, as this case falls between the nearly two-dimensional case and the case of deep aquifers covered by the spectral theory. This intermediate class of flow systems is characterized by aquifer thickness in the range $3m \ll L_3 \ll 100m$ for the case at hand.

For deep aquifers on the other hand, the hydraulic head fluctuates many times in the vertical direction. The infinite domain spectral theory holds in this case, and predicts that the flow system behaves one-dimensionally at the large scale, with the effective conductivity in the mean flow direction equal to the arithmetic mean conductivity. This result indicates that the vertical fluctuations of the conductivity and of the longitudinal head gradient are effectively decoupled. Indeed, the same result could be obtained by vertically averaging the Darcy equation:

$$Q_1 = - K \frac{\partial H}{\partial x_1}$$

and assuming that the fluctuations of K and $\partial H/\partial x_1$ are independent of each other in the vertical.

The limitations of the infinite domain approach for the case of aquifers of finite thickness have been recognized in the past. Naff and Vecchia (1986) developed quasi-analytical

solutions for confined stratified aquifers with finite thickness. They found that the head variance increases to infinity as the ratio of layer thickness to aquifer thickness (ℓ_3/L_3) increases. This is in accordance with the fact that the flow system becomes two dimensional (in this case a Hole-Spectrum is needed in order to obtain finite head variance -- see Table 3.2). However, these authors do not seem to recognize the fact that the vertical head correlation in an infinite domain is large (7ℓ) and nearly independent of the anisotropy ratio. This finding provides a simple criterion for evaluating the range of aquifer thickness for which the infinite domain theory applies: aquifer thickness must be on the order of several tens of horizontal conductivity scales or more. This simple rule does not seem to have been recognized in the past. On the other hand, for very shallow aquifers with thickness on the same order as the horizontal conductivity scale, a two-dimensional theory based on vertical averages could perhaps be used as a first approximation.

Let us now focus exclusively on the case of *deep* stratified aquifers for which the spectral theory holds. In these cases, the hydraulic head standard deviation appears to be proportional to the geometric mean of horizontal and vertical correlation scales ($\sqrt{\ell_1 \ell_3}$), independently of the anisotropy ratio (ℓ_3/ℓ_1). Thus, two different types of aquifers such as (ℓ_1 = 1m, ℓ_3 = 0.2m) and (ℓ_1 = 4m, ℓ_3 = 0.05m) lead to the same head

variance — although the detailed structure of the flow field may be very different. This shows that the value of head variance does not indicate the degree of anisotropy of the velocity field in stratified aquifers.

It is particularly instructive to examine the limit cases $\ell_1, \ell_2 \to \infty$ and $\ell_3 \to 0$ (the anisotropy ratio goes to zero in both cases: limit of perfect stratification). The first case, $\ell_1 \to \infty$, corresponds to an infinite horizontal correlation scale. We have seen that the vertical head correlation length is proportional to ℓ_1, so that it must also become infinite. Therefore, the spectral theory will not hold in this case for any finite aquifer thickness, however large. One must dismiss the case $\ell_1 \to \infty$ as pathological. Examine now the case $\ell_3 \to 0$ corresponding to infinitely small layer thickness. In this case the infinite domain theory seems to apply, and equation (3.3.1) shows that:

$$\begin{aligned} &\sigma_h && \to 0 \\ &\sigma_{q_2},\ \sigma_{q_3} && \to 0 \\ &\sigma_{q_1} && = \sigma_f K_G J_1 . \end{aligned} \tag{3.34}$$

Note that σ_{q_2} and σ_{q_3} vanish, implying that the transverse flux components vanish identically, and the head remains constant in vertical planes across the mean flow direction. Thus the flow

field appears inherently one-dimensional, which explains why the ensemble head variance vanish (consider the equivalence of ensemble mean and 3D averages: the 3D averaged variance of a 1D process must vanish). However, this result turns out to be inconsistent, as mass balance would require that Q_1 be constant for one-dimensional flow. This clearly contradicts the result ($\sigma_{q_1} \neq 0$) obtained above. Again, this case is a singular limit of the three-dimensional spectral theory and must be dismissed.

In summary, we have shown that the limit cases $\ell_1 = \infty$ and $\ell_3 = 0$ are meaningless in the framework of the three-dimensional, infinite domain theory. Nevertheless, the asymptotic analysis of strongly stratified aquifers ($\epsilon \ll 1$) remains valid as long as $\epsilon \neq 0$. The asymptotic solutions (3.31) hold for small anisotropy ratios, provided that the aquifer thickness be "significantly larger" than about 10 ℓ_1, where ℓ_1 is the horizontal conductivity scale. Concerning the flux vector, it is instructive to note that the longitudinal flux component has a variance about twice as large as for the isotropic case, and is independent of ϵ for ϵ small. The variance of the transverse flux components is much smaller, on the order of ϵ, so that nearly one-dimensional flow obtains at the large scale for deep stratified aquifers.

It may be useful to end this section with some remarks concerning the meaning of the so-called "anisotropy ratio". All along, we assumed implicitly that the length scale ratio $\epsilon = \ell_3/\ell_1$ expresses to some kind of anisotropy in the conductive properties of the statistically layered formation. We will show in fact that ϵ is equivalent to the square root of some kind of conductivity anisotropy ratio to be defined shortly.

The interpretation of ϵ in terms of a conductivity anisotropy ratio is obtained by re-scaling the coordinate system in such a way that the random conductivity field becomes isotropic in the rescaled coordinates:

$$x_i' = x_i/\ell_i. \tag{3.35}$$

Starting with an ellipsoidal log-conductivity field $F(\underline{x})$ with covariance function:

$$R_{FF}(\underline{\xi}) = R_{FF}(\sqrt{\xi_1^2/\ell_1^2 + \xi_2^2/\ell_2^2 + \xi_3^2/\ell_3^2})$$

we obtain indeed an isotropic field $F(\underline{x}')$ in the new coordinates:

$$R_{FF}(\underline{\xi}') = R_{FF}(\sqrt{\xi_1'^2 + \xi_1'^2 + \xi_3'^2})$$

Furthermore, the governing flow equation (3.3) written in the rescaled coordinates takes the form:

$$\frac{\partial}{\partial x_i'} \left(K_{ij}(\underline{x}') \cdot \frac{\partial H}{\partial x_j'} \right) = 0 \qquad (3.36)$$

where the $K_{ij}(\underline{x}')$ tensor appears as the product of a deterministic anisotropic conductance tensor and a statistically isotropic random field conductivity, as shown below:

$$K_{ij}(\underline{x}') = \begin{bmatrix} \frac{K_G}{K_{11}} & 0 & 0 \\ 0 & \frac{K_G}{K_{22}} & 0 \\ 0 & 0 & \frac{K_G}{K_{33}} \end{bmatrix} \cdot K(\underline{x}') \qquad (3.37)$$

$$K_{11} = \left[\frac{\ell_1}{\ell_3} \right]^2 K_G$$

$$K_{22} = \left[\frac{\ell_2}{\ell_3} \right]^2 K_G$$

$$K_{33} = K_G$$

In the case of isotropy in the plane of stratification ($\ell_1 = \ell_2$), this yields a formal equivalence between length scale

anisotropy and local deterministic conductivity anisotropy in the form:

$$\epsilon = \frac{\ell_3}{\ell_1} = \sqrt{\frac{K_{33}}{K_{11}}} \qquad (3.38)$$

This suggests that a typical "deterministic anisotropy" on the order $K_{33}/K_{11} \simeq 1/100$, corresponds to a "length scale anisotropy" on the order $\epsilon = 1/10$. However, this interpretation of anisotropy should be taken "with a grain of salt": the *deterministic* anisotropy defined above should be distinguished from the concept of a large scale *effective* anisotropy. Indeed, the spectral solutions developed by Gelhar and Axness (1983) show that the effective conductivity anisotropy is in fact independent from ϵ when ϵ is small, being asymptotically equal to the ratio of harmonic to geometric means. In any case, it is still instructive to think of the length scale ratio (ℓ_3/ℓ_1) as equivalent to the square root of some deterministic conductivity ratio. The form of the scaled flow equation (3.36) also suggests more generally that water flow in a stratified heterogeneous formation results from complex local interactions between purely isotropic random effects and deterministic anisotropy effects.

3.6 Head Moments for the Hole-Markov Spectrum and Low Wavenumber Effects

In order to ascertain that the previous results are physically meaningful, we investigate in this section the behaviour of the head process for a different input spectrum: the Hole-Markov spectrum Table 3.1. The effect of the "hole" here is to reduce the low-wavenumber content of the log-conductivity field, compared to the Markov model used previously. This may affect the behaviour of the head field, particularly by reducing its correlation range.

Assuming again $\ell_1=\ell_1=\ell$ and $\epsilon=\ell_3/\ell$, the 3D anisotropic Hole-Markov spectrum for ℓn K can be written as:

$$S_{ff}(\underline{k}) = \frac{4\sigma_f^2\ell^2\ell^3}{3\pi^2} \cdot \frac{(\ell^2\ (k_1^2+k_2^2) + \ell_3^2k_3^2)}{[1+\ell^2\ (k_1^2+k_2^2) + \ell_3^2k_3^2]^3}$$

Plugging this into (3.10) gives the head covariance in the form

$$R_{hh}(\underline{\xi}) = (\frac{\pi}{8}\sigma_f^2\ J_1^2\ \ell\ell_3) \cdot I(\underline{\xi}) \qquad (3.39)$$

$$I(\underline{\xi}) = (2/\pi\ell)^3 . \iiint_{-\infty}^{+\infty} \frac{k_1^2\ (k_1^2 + k_2^2 + \epsilon^2k_3^2)\cdot\cos(\underline{k}\underline{\xi})}{k^4\ [\ell^{-2}+(k_1^2+k_2^2) + \epsilon^2k_3^2]^3} d\underline{k}$$

The head variance obtains by evaluating the $I(\underline{\xi})$ integral analytically at $\underline{\xi} = \underline{0}$. The computation is detailed in Appendix 3.G, and the result is:

$$\sigma_h^2 = \frac{1}{16} \left(\frac{\pi}{2}\right)^6 \sigma_f^2 J^2 \ell_G^2 \cdot G(\epsilon) \qquad (3.40)$$

with:

$$G(\epsilon) = \frac{1}{\epsilon} \cdot \frac{\epsilon^2}{\epsilon^2-1} \left\{ -1 + \frac{2\epsilon^2-1}{\epsilon\sqrt{\epsilon^2-1}} \ln\left[\frac{\epsilon+\sqrt{\epsilon^2-1}}{\epsilon-\sqrt{\epsilon^2-1}} \right] \right\} \quad \text{for } \epsilon > 1$$

$$G(\epsilon) = 4/3 \qquad \text{for } \epsilon = 1$$

$$G(\epsilon) = \frac{1}{\sqrt{1-\epsilon^2}} \cdot \left\{ \frac{\epsilon}{\sqrt{1-\epsilon^2}} + \frac{1-2\epsilon^2}{1-\epsilon^2} \cdot \arcsin(\sqrt{1-\epsilon^2}) \right\} \quad \text{for } \epsilon < 1.$$

When the geometric scale $\ell_G = \sqrt{\ell_1 \ell_3}$ is kept fixed, this gives the following asymptotic results for the strongly anisotropic cases $\epsilon \gg 1$ (vertical layers) and $\epsilon \ll 1$ (horizontal layers):

$$\epsilon \to \infty: \quad \sigma_h^2 \approx \frac{1}{16} \left(\frac{\pi}{2}\right)^6 \sigma_f^2 J^2 \ell_G^2 \cdot \frac{4\ln(2\epsilon)-1}{\epsilon} \to 0 \qquad (3.41)$$

$$\epsilon \to 0: \quad \sigma_h^2 \approx \frac{1}{16} \left(\frac{\pi}{2}\right)^7 \sigma_f^2 J^2 \ell_G^2 \qquad (3.42)$$

In the isotropic case ($\epsilon = 1$), the head variance is:

$$\epsilon = 1: \quad \sigma_h^2 = \frac{1}{12} \left(\frac{\pi}{2}\right)^6 \sigma_f^2 \, J^2 \, \ell_G^2 \tag{3.43}$$

The results obtained for the isotropic case ($\ell_3=\ell_1$) and the anisotropic case ($\ell_3<<\ell_1$) resemble those previously obtained with the Markov spectrum without a hole. The following table summarizes the different values of σ_H obtained for the Markov and Hole-Markov spectra for various degrees of anisotropy.

Anisotropy $\epsilon=\ell_3/\ell_1$	Markov Spectrum (λ_i): $\sigma_h/(\sigma_f \, J\lambda_G)$	Hole-Markov Spectrum (ℓ_i): $\sigma_h/(\sigma_f \, J\ell_G)$
$\epsilon << 1$:	$\sqrt{\pi/8} \approx 0.63$	$(\pi/2)^{7/2}/4 \approx 1.21$
$\epsilon = 1$:	$1/\sqrt{3} \approx 0.58$	$\sqrt{1/12}\,(\pi/2)^3 \approx 1.12$
$\epsilon >> 1$:	-----	$\frac{4\ln(2\epsilon)}{\epsilon} \to 0$

The comparative table above shows that the head standard deviations will not be the same if the ℓ_i-scales of the Hole-Markov are taken equal to the λ_i-scales of the Markov spectrum. Remarkably, it appears that the same head variances are obtained with the two spectra by taking the length scales of the Hole-Markov spectrum to be half the correlation scales of the Markov spectrum, i.e.:

$$\ell_1 = \lambda_1/2.$$

Furthermore, it appears that the head variance for strongly anisotropic formations (such that $\ell_3 \ll \ell_1$) is approximately equal to the head variance obtained for an equivalent isotropic medium with correlation scale $\ell_G = \sqrt{\ell_1 \ell_3}$. Recall that the same observation *holds for the Markov model*. In addition, the table also shows that the head variance vanishes in the limit for strongly anisotropic *vertical* formations ($\ell_3 \gg \ell_1$) if ℓ_G is kept constant.

In order to complete the comparison between the Markov and Hole-Markov spectral models, it may be instructive to compare the head correlation functions obtained in the two cases. This would help quantify the influence of the input log-conductivity spectrum (or correlation function) upon the spatial structure of the head field. Specifically here, one may expect that the "Hole" model (spectrum with a low wavenumber hole) produces a head field with smaller correlation length. The question is: how important is this effect? And finally: how sensitive is the spatial structure of the solution with respect to the assumed shape of the log-conductivity spectrum?

For the Hole-Markov model, Appendix (3.G) develops a

close-form expression for the transverse correlation function $R_{hh}(0,0,\xi)$ in the isotropic case $\epsilon=1$. The final result is reproduced below:

$$\epsilon = 1: \frac{R_{hh}(0,0,\xi_3)}{\sigma_h^2} = \frac{12}{s_3}\left\{\frac{2}{s_3}\left[\frac{1-e^{-s_3}}{s_3} - e^{-s_3}\right] - (1 + \frac{s_3}{4})e^{-s_3}\right\} \tag{3.44}$$

One way to compare the Markov and Hole-Markov models is to evaluate the e-correlation lengths, i.e., separation distance at which the correlation drops below e^{-1}. The e-correlation length of head is about 7λ for the Markov-model, and about 3ℓ for the Hole-Markov model. This indicates that the head correlation across flow is about half smaller for the Hole model, based on $\lambda_i=\ell_i$. On the other hand, recall that the head variances obtained for the two models become equal if one chooses $\ell_i=\lambda_i/2$. With this choice, the head correlation would appear 4 times smaller when using the Hole model. Perhaps the most rational approach is to choose the scales ℓ and λ in such a way that the e-correlation scales of the two random fields coincide. It can be shown that this particular choice corresponds to $\ell = \lambda/0.723$ (from Table 2.2 of Vomvoris, 1986). It does not seem that this choice leads to a better agreement between the head solutions obtained by the two models.

We conclude that the spatial structure of the head field is fairly sensitive to the choice of the log-conductivity spectrum. For the 3D isotropic case, the head correlation lengths are increased with increasing spectral density at low wavenumbers, all other quantities being held constant (fixed head variance). On the other hand, remember that the head random field becomes non-stationary with infinite correlation length in two dimensions, unless the spectral density of log-conductivity vanishes at zero wavenumber (see Table 3.2). This indicates that the effect of large scale conductivity fluctuations becomes more significant as the degree of freedom of flow decreases from 3 to 2-dimensional flow. One may think of 3D anisotropy as an intermediate case between the 2 and 3-dimensional isotropic cases.

Our finding that the large scale fluctuations of the conductivity field have a significant effect on the spatial structure of the head field raises new questions about the appropriate determination of conductivity spectra from field measurements. In our view, the available spectral analysis of field-measured conductivities (Bakr, 1976) do not lead to favor the Hole-Markov over the Markov-spectrum or vice-versa. This is

due to the fact that the low wavenumber range of the spectrum cannot be determined with reasonable confidence for wavenumbers near or below the inverse domain size (standard estimates of spectral confidence interval also break down in this range). Rather, we feel that the size of the domain of interest should be used as an extra parameter to determine the appropriate cut-off of the "measured" log-conductivity spectrum at some low wavenumber, in such a way that essentially all fluctuations larger than the inverse domain size be removed. The Hole-Markov model is just one way of carrying on the cut-off procedure in an implicit way. However, a different approach, which departs from the "infinite domain" postulate, will be developed in Section 4.4 of Chapter 4 in order to clarify the effects of finite domain size.

CHAPTER 4: EXTENSIONS OF SPECTRAL THEORY: NON-PERTURBATIVE SOLUTIONS, SPECTRAL CONDITIONING AND UNCERTAINTY

4.1 Introduction: Sources of Errors in Standard Spectral Solutions

This chapter is devoted to the improvement and generalization of the spectral perturbation solutions of saturated flow developed earlier (Chapter 3). By exploring the stochastic flow problem from a somewhat broader viewpoint, we hope to shed some light on the approximations involved in the standard spectral solutions (see also Chapter 6 for a comparison with direct numerical simulations). More importantly, our goal is to develop alternative approaches for obtaining realistic yet tractable solutions of stochastic flow, and related phenomena like dispersive solute transport.

One of the major advantages of the spectral perturbation theory, as it stands, is its high potential for producing tractable closed form results. On the other hand, this theory presumably suffers some drawbacks due to the approximations that were made. In the forthcoming sections, we will develop *non-perturbative* solutions method (Section 4.2), as well as new perturbative solutions (Section 4.3), and suggest possible extensions of the spectral theory to include non-stationarity or finite-size effects (Section 4.4). In the preliminary study that

follows, we review some of the most significant sources of "errors" in order to clarify the weak points of the current spectral theory.

There are two levels of approximations involved in the standard spectral solution method: first, the "*solution errors*" due to the approximate solution of the postulated stochastic flow equation, and second, the "*model errors*" due to the approximate representation of real world heterogeneities by stationary and ergodic random fields. The perturbation approximations belong to the first category, while the basic assumptions of infinite domain, statistical homogeneity, and ergodicity of the log-conductivity field, belong to the second category. Let us now review in some detail the potential inaccuracies of the current spectral theory, due to approximations made at the "solution" level and at the "model" level, respectively.

Solution errors can be identified through a formal, qualitative comparison of spectral and exact solutions, assuming that the basic premises of the spectral theory are true. Hence, in keeping with the spectral approach, let us assume for the moment that the flow field is indeed governed by the stochastic equation (3.1), that the domain is infinite, and the log-conductivity $\ln K(\underline{x})$ is a Gaussian, stationary and ergodic random field, with first and second order moments invariant by

translation. Based on these premises, the standard spectral perturbation approach (Chapter 3) requires some additional hypotheses and approximations in order to arrive at closed form solutions:

(i) All second order and higher order terms $O(\sigma^p)$, $p \geq 2$, are neglected in the flow equation governing head.

(ii) The random fields h, $\underline{\nabla}H$ and $\underline{Q}$ are assumed to be stationary and ergodic in the first and second moments (recall that h is the head perturbation $H - \langle H \rangle$).

(iii) The random fields h, $\underline{\nabla}H$ and $\underline{Q}$ are implicitly assumed to be nearly Gaussian, so that knowledge of their first and second moments (mean and covariance function) suffices to determine their statistical properties entirely.

Intuitively, the validity of all three requirements depends on σ_f being small. This is particularly obvious for (i), but not as obvious for (ii) and (iii). Some results in the literature suggest that stationarity and normality (for heads) are satisfied asymptotically as $\sigma_f \to 0$, but may not hold as σ_f increases. Specifically, Gutjahr and Gelhar (1981) showed that if $\sigma_f \ll 1$ then the head field is stationary in the case of three-dimensional isotropic $\ln K$ fields whose covariance function satisfies:

$$\int_0^\infty \xi \, R_{ff}(\xi) d\xi < \infty.$$

This relation is satisfied in particular for any isotropic $R_{ff}(\xi)$ that is positive everywhere and has finite integral correlation length (see Table 2, Chapter 3). Also by looking at the head covariance (3.22) obtained for a Markov spectrum, one can see that:

$$\lim_{\xi \to \infty} R_{hh}(\xi) = 0$$

which is a sufficient condition for ergodicity in the first and second order increments (cf. Yaglom, 1962, 1.4). Unfortunately, this only proves that $h(\underline{x})$ is *asymptotically* stationary and ergodic as $\sigma_f \to 0$ — since a perturbation method was used to establish the proof. In fact, it appears from the detailed perturbation analysis (3.4-3.6) that the apparent stationarity in the mean head gradient results from neglecting second order terms $0(\sigma^2)$ in the equations. This suggests that the solution may not be stationary unless σ_f is small.

Similarly, Gutjahr (1984) showed that the first order spectral perturbation solutions are exact for arbitrary σ_f *if it can be assumed that* $(\underline{y}H, f)$ *are jointly Gaussian.* However, the point is precisely that the random fields $\underline{y}H$ and $\underline{Q}$ become increasingly skewed (non-Gaussian) as σ_f increases. In turn,

this implies that the standard spectral perturbation method becomes increasingly inaccurate as the skewness of the solution increases with σ_f. Indeed, non-Gaussian random fields cannot be completely characterized with first and second moments alone. Some of the numerical experiments of Chapter 6 will confirm the non-Gaussian character of the flux or velocity field.

In summary, we argue that a large log-conductivity variance could yield non-negligible high order terms in the perturbation equations. This in turn produces two types of effects that cannot be captured by the first order spectral perturbation solution:

(i) non stationary behavior of first and second moments

(ii) non Gaussian distribution of the random field solution.

At first sight, it seems natural to try developing higher order perturbation expansions in order to predict more accurately the statistical behavior of highly variable flow fields. An effort in this direction was pursued by Dagan (1985). However, this author did not obtain third or higher order moments, as would be needed to characterize a non-Gaussian behavior. In addition, it should be kept in mind that the perturbation expansion admittedly may not converge at all for $\sigma_f > 0$, even though the first order term does converge to the

exact solution as $\sigma_f \to 0$.

The distinction between convergent and asymptotic expansion is well documented in the literature (see Bender and Orszag, 3.8, 1978). Consider for instance the following expansion of the head perturbation h in the small parameter σ:

$$h^{(N)} = h_0 + \sigma h_1 + \sigma^2 h_2 + \cdots + \sigma^N h_N \qquad (4.1)$$

The Nth-order approximation $h^{(N)}$ may possibly be asymptotic to the exact solution h, i.e.:

$$|h^{(N)} - h| << \sigma^N \text{ as } \sigma \to 0, \text{ N fixed}$$

... even though the series diverges, i.e.:

$$\lim_{N \to \infty} |h^{(N)} - h| \neq 0, \ \sigma > 0 \text{ fixed.}$$

Accordingly, one should not expect too much improvement from higher order solutions beyond the first few terms; there is even the possibility that "higher order" approximations be less accurate for a given, fixed value of σ. In fact, the work by Dagan (1985) shows that the head covariance is not changed much by using second order rather than first order expansions. For instance, the head variance obtained by numerical integration of

Dagan's spectral solution for a 3D isotropic Markov field is only slightly decreased as follows (Lynn Gelhar, personal communication):

$$\mathrm{Var}(h)_2 = \frac{1}{3}\,(1-0.058\,\sigma_f^2)\,\sigma_f^2 J^2\lambda^2 \qquad (4.2)$$

compared to the first order result (Eq. 3.21):

$$\mathrm{Var}(h)_1 = \frac{1}{3}\,\sigma_f^2 J^2\lambda^2.$$

The difference between these two expressions is quite mild for log-conductivity standard deviations on the order of unity. This indicates that the head variance is relatively unaffected by high order interactions. However, the most interesting high order effects may have been "missed", as only three-point covariance functions can capture the "skewness effects" due to large σ_f. Moreover, it is also worth noting that the type of higher order spectral perturbation such as used by Dagan (1985) does not address the "large variance nonstationarity effects" discussed above.

In the present work, we will not pursue the classical approach of developing higher order expansions. Rather we develop a more general non-perturbative approach based on exact statistical identities in order to assess the validity of the

standard first order spectral solutions (Section 4.2). This will lead us also to propose a new perturbation solution for the flux spectrum (Section 4.3).

Let us now focus briefly on the second category of errors, the so-called "*model errors*" defined earlier. In view of real field situations, it would seem that it is not always possible to identify uniquely a stationary $\ln K$ field with a definite correlation scale. Thus, the ideal case postulated in Chapter 3 may not be encountered often in actual practice. In our view, this type of identification problem may be due in practice to inadequate sampling of field data (geometry and spacing of measurement network), and/or to the particular spatial structure of the subsurface formation, which may involve some large scale inhomogeneities. This kind of situation could lead to inconsistent random field identification, for instance with apparent correlation scales on the order of domain size (see some of the results reported by Hoeksema and Kitanidis, 1985). With proper detrending, however, it is often possible to identify a definite correlation length based on the assumption that the detrended field is stationary. Nevertheless, the correlation scale determined in this manner may still depend on the particular subregion of investigation, as suggested by some field studies. The reader is referred to the data review of Chapter 2 for more details and references.

Another related situation of interest is the case where the flow or transport process of interest takes place on increasingly larger regions as time evolves. This occurs in a number of cases of great importance for contamination studies, for instance in the case of a contaminant plume spreading from a local source in an aquifer, or an unsaturated moisture plume spreading from a local infiltration area. In both cases, the "*global scale*" of interest evolves in time, and it is conceivable that the most dominant contributions from spatial heterogeneities occur at varying length scales, as the size of the plume evolves. This idea could be related to our earlier observation that the formation's heterogeneity is only *locally* stationary around some given trend, and with a given correlation scale, both depending on the size of the region. This type of finite scale problem will be approached in the last section of this chapter (Section 4.4) by using band-pass self-similar spectra, and by developing the idea of *spectral conditioning*. The preliminary results obtained there will show explicitly the scale dependence and uncertainty of the head variance, effective conductivity, and macrodispersivity for flow and transport phenomena taking place over finite domains.

4.2 Non-Perturbative Spectral Solutions and Statistical Symmetries

4.2.1 Summary:

In this lengthy section, we show that a *stationary* solution of the stochastic flow equation, if it exists, must satisfy a certain set of statistical relations, notably in terms of the spectrum of the flux vector and head gradient. These statistical identities are derived directly from the continuity equation, and by taking into account the inherent symmetries of the flow system in any number of dimensions, particularly in the case of 3D and 2D isotropic media. These relations are used to "test" the standard spectral solutions. In the special 2D isotropic case, the effective conductivity must be identical to the geometric mean, and an exact relation is found between the flux spectrum and head spectrum. Both results are based on a conjugacy property relating the flux and head gradient in two-dimensional space with isotropic Gaussian log-conductivity. The problem of determining a general relation for the effective conductivity tensor in 3D anisotropic media is also investigated, leading to a general closed form relation in terms of the log-conductivity variance and two anisotropy length scale ratios. Note that the spectral relation obtained in the 2D case also suggested a modification of the standard spectral solutions, to be developed in a forthcoming section (4.3).

4.2.2 Mass Conservation Relation:

The mass conservation or "continuity" equation for steady flow in m-dimensional space is:

$$\frac{\partial Q_i}{\partial x_i} = 0 \qquad (i = 1,\cdots,m) \qquad (4.3)$$

where the implicit Einstein summation convention was used. We now assume that the basic premises of the standard spectral theory hold, i.e., the flux Q_i is a *stationary* random vector field. Whence the mean $\bar{Q}_i = \langle Q_i(x) \rangle$ is a constant vector, and the perturbation $q_i(\underline{x}) = Q_i(\underline{x}) - \bar{Q}_i$ is zero mean stationary. For all practical purposes here, second order stationarity suffices.

By averaging the continuity equation (4.3) and then substracting, one obtains an equation for the mean and another equation for the perturbation, of identical form:

$$\frac{\partial \bar{Q}_i}{\partial x_i} = 0 \qquad (i = 1,\cdots,m) \qquad (4.4)$$

$$\frac{\partial q_i}{\partial x_i} = 0 \qquad (i = 1,\cdots,m). \qquad (4.5)$$

Because these equations are *stochastically linear*, and the flux

Q_i was assumed stationary, the spectral representation theorem (3.7) can be used to obtain an exact relation on the spectrum of q_i from equation (4.5). The Fourier increments satisfy:

$$j\, k_i\, dZ_{q_i}(\underline{k}) = 0. \qquad (j = \sqrt{-1}) \tag{4.6}$$

Multiplying by $dZ^*_{q_j}(\underline{k})$ and averaging yields:

$$k_i \cdot S_{q_i q_j}(\underline{k}) = 0 \qquad (j = 1 \cdots m) \tag{4.7}$$

where $S_{q_i q_j}$ is the tensor spectrum of the flux vector, whose Fourier Transform is the tensor covariance function $R_{q_i q_j}(\underline{\xi})$. The equivalent mass conservation condition on $R_{q_i q_j}(\underline{\xi})$ is easily obtained by Fourier-transforming (4.7):

$$\frac{\partial R q_i q_j(\underline{\xi})}{\partial \xi_i} = 0 \qquad (j = 1 \cdots m). \tag{4.8}$$

Now, it is easily seen that the standard first order spectral solutions developed in Chapter 3 verify mass conservation both in the mean and second order moments. The mean equation (4.4) is automatically satisfied since $\overline{Q}_i$ is constant; and the equation (4.7) for $S_{q_i q_j}$ is satisfied by the spectrum given in (3.17), as can be easily verified. We conclude that the

standard spectral solutions are self-consistent with respect to mass conservation, as far as first and second moments are concerned. Incidentally, we obtain for the special 1D case:

$$k\, S_{qq}(k) = 0$$

which shows that $S_{qq}(k) = 0$ is a solution, i.e., *the flux* q *must be a deterministic constant*. It may seem that a flux spectrum of the form $S_{qq}(k) = \sigma_q^2\, \delta(k)$ could also be solution, i.e., the flux q could be a spatially constant random variable with variance σ_q^2. However this gives rise to an indeterminacy in the limit $kS_{qq}(k)$ as $k \to 0$ (this can be seen by replacing $\delta(k)$ by any sequence of functions that converges to $\delta(k)$). Therefore, it seems that the case $\sigma_q \neq 0$ should not be accepted as a valid "stationary" solution in the 1D case. Gutjahr and Gelhar (1981) adopted a different view in their discussion of stationary and non-stationary two-point boundary value flow problems with $\sigma_q \neq 0$.

Finally, let us point out a result mentioned in Batchelor (1953). By using the continuity condition (4.7) along with the usual properties of a spectral density tensor:

$$S_{q_i q_j}(\underline{k}) = S_{q_j q_i}(-\underline{k}) \qquad (4.9)$$

he obtained the following spectrum:

$$S_{q_i q_j}(\underline{k}) = \beta^2(\underline{k}) \cdot \left[\delta_{ij} - \frac{k_i k_j}{k^2}\right] + \alpha_i(\underline{k})\alpha_j^*(\underline{k}) \left[1 - \frac{\beta^2(\underline{k})}{\alpha^2(\underline{k})}\right]. \quad (4.10)$$

Here, the complex vectors $\alpha_i(\underline{k})$, $\beta_i(\underline{k})$ are mutually orthogonal, and orthogonal with the wavenumber vector (i.e., $\underline{\alpha}\cdot\underline{\beta} = \underline{\alpha}\cdot\underline{k} = \underline{\beta}\cdot\underline{k} = 0$). The scalar quantities α^2, β^2, k^2 denote the squared modulus of the vectors $\underline{\alpha}$, $\underline{\beta}$, $\underline{k}$, respectively. We will see that one can arrive at more useful special forms of the flux spectrum by considering, along with the continuity equation, the invariance of the tensor under certain transformations arising from spatial symmetries of the random flow problem, particularly for statistically isotropic conductivities in 2D and 3D space.

4.2.3 Statistical Axial Symmetry for 3D Flow in Isotropic Media:

Here we consider the case of flow in a statistically isotropic formation in the infinite 3D space. The log-conductivity F= ln K is a statistically isotropic homogeneous Gaussian random field, implying in particular that the covariance function $R_{ff}(\underline{\xi})$ depends only on the radial separation distance $\xi = \sqrt{\xi_1^2+\xi_2^2+\xi_3^2}$ (similarly the spectrum $S_{ff}(\underline{k})$ depends only on $k = \sqrt{k_1^2+k_2^2+k_3^2}$). We now examine the consequences of this symmetry, assuming as before that the solution (h, $\underline{y}$H, and $\underline{q}$) is

statistically homogeneous in space. In particular, we seek relations on the flux covariance $R_{q_i q_j}(\underline{\xi})$ or on the spectral tensor $S_{q_i q_j}(\underline{k})$, in a manner somewhat similar to previous work on the statistical theory of isotropic turbulence (see the books by Batchelor 1953, Monin and Yaglom 1965, and Landahl and Mollo-Christensen 1986). However, note that the case at hand is different from isotropic turbulence in one important respect: the flow driven by a mean gradient in a random porous medium cannot be represented realistically as a statistically isotropic field in all three space dimensions, even when the underlying medium is fully isotropic. Rather, the flow field can be thought of as having statistical *axial symmetry* with respect to the mean flow direction.

In order to clarify this notion, we turn back to the basic flow equations expressed in terms of "detrended" random fields. The perturbation equations for the head and the flux vector are, respectively:

$$\frac{\partial^2 h}{\partial x_i \partial x_i} + \frac{\partial f}{\partial x_i}\frac{\partial h}{\partial x_i} - \bar{J}_1 \frac{\partial f}{\partial x_1} = 0$$

$$\bar{Q}_i = K_G J_1 \delta_{1i}\, e^{\sigma^2/2} - K_G \langle e^f \frac{\partial h}{\partial x_i} \rangle \qquad (4.11)$$

$$q_i = K_G\,(e^f - e^{\sigma^2/2})(J_1\delta_{1i} - \frac{\partial h}{\partial x_i}) - K_G\, e^{\sigma^2/2}\frac{\partial h}{\partial x_i} - (\bar{Q}_i - K_G e^{\sigma^2/2})$$

where we used the stochastic Darcy equation for obtaining $\bar{Q}_i$ and

q_i. Recall that h and q_i are the statistically homogeneous perturbations of the head and the flux vector, $\bar{J}_1$ is the only non-zero component of the constant mean hydraulic gradient (by choice of the x_1 axis), and $\bar{Q}_i$ is the i-th component of the mean flux vector. Recall also that the log-conductivity perturbation $f(\underline{x})$ is Gaussian, and $R_{ff}(\underline{\xi})$ is invariant under translations and rotations in the $\underline{\xi}$-space. In addition, all higher order moments of $f(\underline{x})$ are expressible in terms the covariance of R_{ff}, and the N-point covariance $\langle f(\underline{x}_1)\cdots f(\underline{x}_N)\rangle$ is also invariant under translations and rotations in $\underline{x}$-space. The form taken by (4.11) implies that $h(\underline{x})$ must then be statistically invariant under rotations in the (x_2,x_3) plane, as well as invariant under reflections through the x_1 axis. In particular $R_{hh}(\underline{\xi})$ is isotropic in (ξ_2,ξ_3), and invariant when (ξ_1) is changed into $(-\xi_1)$. In addition, we expect that the vectors $g_i = -\frac{\partial h}{\partial x_i}$, and perhaps q_i, be also invariant under such transformations. We now explore this in more detail, beginning with the head gradient vector g_i.

By using the fact that g_i is a special kind of "potential" vector, being the gradient of a scalar quantity $h(\underline{x})$ that is statistically isotropic in the (x_2,x_3) plane, one can obtain explicit relations on the covariance tensor $R_{g_ig_j}(\underline{\xi})$. Indeed, it is well known that, if h is a homogeneous field, then $g_i = \frac{\partial h}{\partial x_i}$ has the covariance tensor:

$$R_{g_i g_j}(\xi) = - \frac{\partial^2 R_{hh}(\xi)}{\partial \xi_i \partial \xi_j}. \qquad (4.12)$$

On the other hand, $R_{hh}(\xi)$ is isotropic in the (ξ_2, ξ_3) plane as explained earlier. Thus we can write:

$$R_{hh}(\xi) = R_{hh}(\xi_1, r) \qquad (4.13)$$

where R is an even function of ξ_1, and r is the 2D radial separation distance $(\xi_2^2 + \xi_3^2)^{1/2}$ in the cross-flow-plane. By applying the chain rule of differentiation, and using the fact that $\partial r/\partial \xi_i = \xi_i/r$ for $i = 2$ and 3, we obtain finally the head gradient covariance in terms of the head covariance:

$$R_{g_1 g_1}(\xi) = - \frac{\partial^2 R_{hh}}{\partial \xi_1^2}$$

$$j \neq 1: \; R_{g_1 g_j}(\xi) = R_{g_j g_1}(\xi) = - \frac{\xi_j}{r} \cdot \frac{\partial^2 R_{hh}}{\partial r \partial \xi_1} \qquad (4.14)$$

$$i \neq 1, \; j \neq 1: \; R_{g_i g_j}(\xi) = R_{g_j g_i}(\xi) = - \frac{\delta_{ij}}{r} \frac{\partial R_{hh}}{\partial r}$$

$$- \frac{\xi_i \xi_j}{r^2} \left[\frac{\partial^2 R_{hh}}{\partial r^2} - \frac{1}{r} \frac{\partial R_{hh}}{\partial r} \right].$$

These equations give the general form of the hydraulic gradient covariance tensor for the case of flow in a 3D isotropic medium. Note that $R_{g_i g_j}$ is a symmetric tensor and an even function of $(\underline{\xi})$ for all (i,j), so that the identity $R_{g_i g_j}(\underline{\xi}) = R_{g_j g_i}(-\underline{\xi})$ is satisfied, as it should for any tensor covariance of a statistically homogeneous vector field. In addition, it can be seen by inspection of (4.14) that $R_{g_i g_j}(\underline{\xi})$ is invariant under the rotation-reflexion group restricted to the transverse flow plane (x_2, x_3). Here, rotational invariance is to be understood in the sense of tensor invariance: the tensor function is invariant when expressed in the new coordinates according to tensor transformation rules. For the case of pure rotations we have:

$$[a_{ij}] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$$

with transformation rules:

$$x'_j = a_{ij}\, x_i; \quad x_i = a_{ji}\, x'_j$$

$$\frac{\partial \bullet}{\partial x_i} = a_{ij}\, \frac{\partial \bullet}{\partial x'_j}; \quad T_{ij} = a_{mi}\, a_{\ell j}\, T'_{m\ell} \qquad (4.15)$$

The same transformation rules apply for pure reflexions, defined

as:

$$[a_{ij}] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Invariance is verified by checking directly from (4.14) that $R_{g_i g_j}$ indeed transforms according to:

$$R_{g_i g_j}(\underline{\xi}) = a m_i a \ell_j \, R'_{g'_m g'_\ell}(\underline{\xi}').$$

For the reflexions in particular, this implies that the two-point covariance of the head gradient is invariant under transformation $\xi_2 \to -\xi_2$ or $\xi_3 \to -\xi_3$. In addition, it can be seen from (4.14) that $R_{g_i g_j}(\underline{\xi})$ is also invariant under reflexions through the x_1 axis ($\xi_1 \to -\xi_1$). Finally, it is worth noting that the cross-covariances vanish along certain lines or planes, as can be seen in Figure 4.1:

$$\begin{aligned} R_{g_1 g_2}(\xi_1, 0, \xi_3) &= 0 \\ R_{g_1 g_3}(\xi_1, \xi_2, 0) &= 0 \\ R_{g_2 g_3}(\xi_1, 0, 0) &= 0 \end{aligned} \qquad (4.16)$$

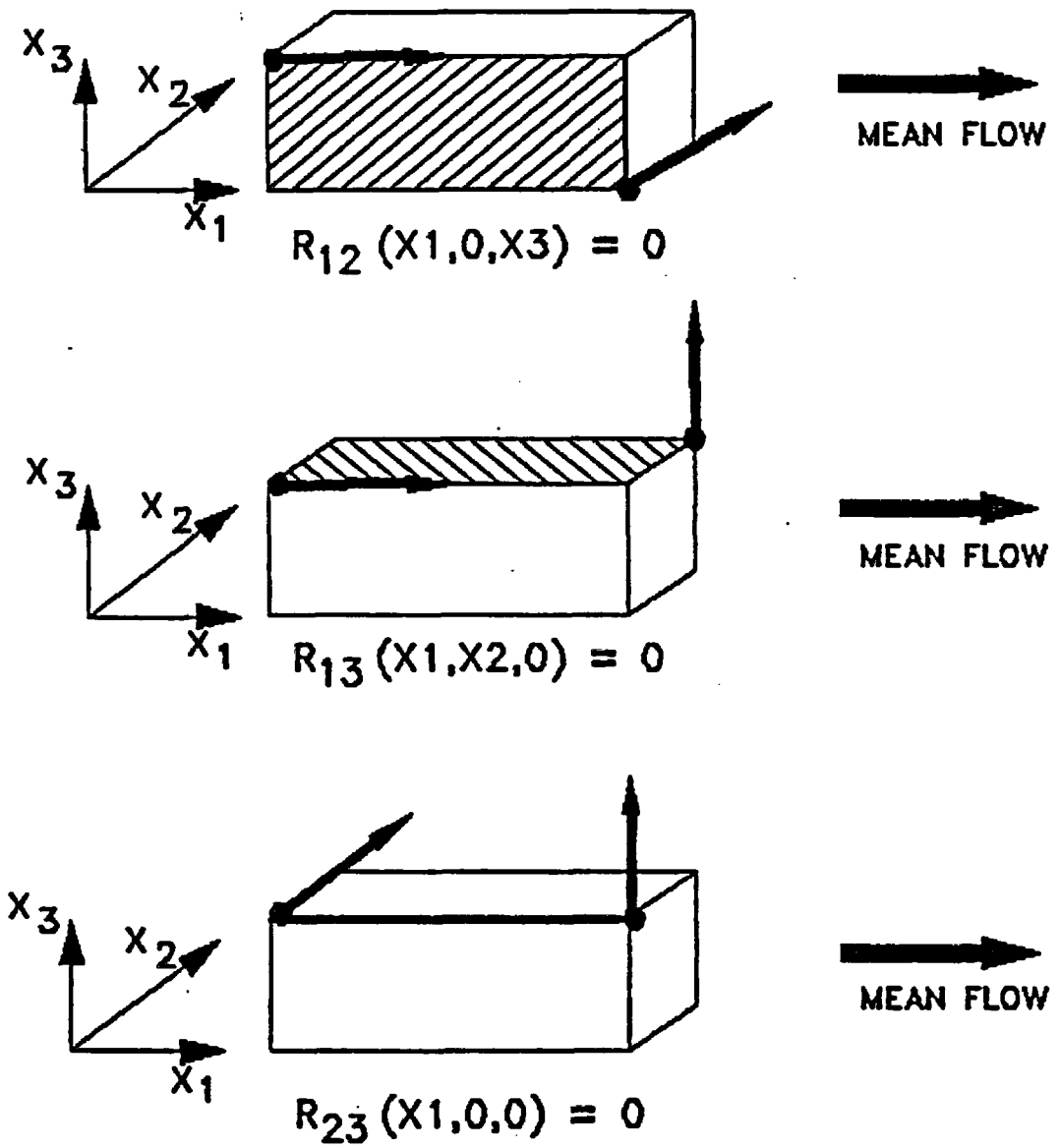


Figure 4.1 The cross-covariance function of the head gradient and flux vectors vanish along certain directions in a 3D isotropic medium.

Note in particular that all the cross-covariances at lag zero ($\underline{\xi} = \underline{0}$) are null. The result in (4.16) can be found by inspection of (4.14), or by looking directly at the consequences of invariance under reflexions and rotations of the $R_{g_i g_j}(\underline{\xi})$.

The results obtained so far can be summarized as follows:

(i) For a 3D isotropic medium, the covariance tensor $R_{g_i g_j}(\underline{\xi})$ of the head gradient is symmetric, invariant under rotations and reflexions in the plane transverse to flow, and under reflexions through the mean flow direction. Invariance is understood to hold under transformations of the coordinate system provided application of the usual tensor transformation rules.

(ii) The general form of $R_{g_i g_j}(\underline{\xi})$ is given by (4.14), and the cross-covariances in particular vanish along certain directions according to (4.16).

It is important to keep in mind that all the symmetry relations developed above are independent of the small variance approximations that were used to obtain first order spectral solutions. We now show that the first-order solutions previously

obtained for R_{hh} and $R_{g_ig_j}$ do satisfy the symmetry properties (4.14, 4.16). This can be easily seen by observing that the $R_{hh}(\underline{\xi})$ given in (3.22) is indeed statistically isotropic in the transverse plane, and so is the spectrum $S_{hh}(\underline{k})$. As a consequence, the covariance and spectrum of $g_i = \partial h/\partial x_i$ obtained from the first order theory satisfy indeed all the invariance properties outlined above. Incidentally, note that the covariance $R_{g_ig_j}(\underline{\xi})$ can be obtained in close form from the first order solution $R_{hh}(\underline{\xi})$ by using (3.22) and (4.14), thus avoiding some difficult Fourier integral.

Finally, it can be shown by using the properties of the Fourier Transform that the spectrum tensor $S_{g_ig_j}(\underline{k})$ shares the same invariance properties as $R_{g_ig_j}(\underline{\xi})$. More precisely, it is easily seen from $g_i = \partial h/\partial x_i$ that $S_{g_ig_j}$ must be of the form:

$$S_{g_ig_j}(\underline{k}) = - k_i k_j \; S_{hh}(\underline{k}) \qquad (4.17)$$

where the spectrum S_{hh} must be invariant under rotations and reflexions in the transverse plane (k_2, k_3) -- and reflexions through the longitudinal direction k_1. All the symmetry properties previously established for $R_{g_ig_j}$ could be deduced from (4.17) by plugging $S_{hh}(\underline{k}) = S(k_1, k)$ and using Fourier Transforms.

We now focus our interest on the flux vector perturbation q_i, which is the physical quantity of interest for applications to contaminant transport. The Darcy equation can be used to decompose q_i into a sum of terms involving the random field perturbations: $p(x) = e^{f(\underline{x})} - e^{\sigma^2/2}$ and $g_i = \partial h/\partial x_i$. Using equation (4.11) and decomposing leads to:

$$q_i(\underline{x}) = q_i^{(1)}(\underline{x}) + q_i^{(2)}(\underline{x})$$

$$q_i^{(1)}(\underline{x}) = - K_G\{p(\underline{x})\, J_1\delta_{1i} - e^{\sigma^2/2}\cdot g_i(\underline{x})\} \qquad (4.18)$$

$$q_i^{(2)}(\underline{x}) = - K_G\{p(\underline{x})\cdot g_i(\underline{x}) - \langle p(\underline{x})\cdot g_i(\underline{x})\rangle\}.$$

The term $K_G\, p(\underline{x})$ is the perturbation of the log-normal conductivity field $K(\underline{x}) = K_G \exp(f(\underline{x}))$, and $g_i(\underline{x})$ is the perturbation of the head gradient vector as before. The first term $q_i^{(1)}$ in (4.18) represents the flux perturbation produced by the separate contributions from the conductivity and head gradient fluctuations while the second term $q_i^{(2)}$ involves the stochastic interactions between them.

The covariance tensor $R_{q_iq_j}(\underline{\xi})$ could be worked out in principle by computing all the terms involved in $\langle q_i(x)\cdot q_j(\underline{x}+\underline{\xi})\rangle$. This leads to an expression involving third and fourth order moments of the augmented vector $(p(\underline{x}), g_i(\underline{x}))$. Assuming that

$(p(\underline{x}).g_i(\underline{x}))$ is jointly statistically homogeneous, up to at least fourth order moments, one obtains triple-point and four-point covariances of the form:

$$\langle u(\underline{x})v(\underline{x}+\underline{\xi}_v)\ w(\underline{x}+\underline{\xi}_w)\rangle = R_{uvw}(\underline{\xi}_v,\underline{\xi}_w)$$

$$\langle u(\underline{x})v(\underline{x}+\underline{\xi}_v)\ w(\underline{x}+\underline{\xi}_w)\ z(\underline{x}+\underline{\xi}_z)\rangle = R_{uvwz}(\underline{\xi}_v,\underline{\xi}_w,\underline{\xi}_z).$$

With these definitions in mind, we obtain from equation (4.18) a general expression for the flux covariance tensor:

$$R_{q_iq_j}(\underline{\xi}) = R_{ij}^{11}(\underline{\xi}) + R_{ij}^{12}(\underline{\xi}) + R_{ij}^{22}(\underline{\xi}) \qquad (4.19.a)$$

where:

$$R_{ij}^{11} = K_G^2 \cdot \{J_i^2\cdot\delta_{1i}\delta_{1j}R_{pp}(\underline{\xi}) + e^{\sigma^2}\cdot R_{g_ig_j}(\underline{\xi}) - e^{\sigma^2/2}J_i(\delta_{1i}R_{pq_j}(\underline{\xi}) + \delta_{1j}R_{pg_i}(-\underline{\xi}))\}$$

$$R_{ij}^{12} = K_G^2 \{J_i\delta_{1i}(R_{ppg_j}(\underline{\xi},\underline{\xi}) + R_{ppg_j}(-\underline{\xi},-\underline{\xi})) - e^{\sigma^2/2}\cdot(R_{pg_ig_j}(\underline{0},\underline{\xi}) + R_{pg_ig_j}(\underline{0},-\underline{\xi}))\} \qquad (4.19.b)$$

$$R_{ij}^{22} = K_G^2 \{R_{pg_ipg_j}(\underline{0},\underline{\xi},\underline{\xi}) - R_{pg_i}(\underline{0})\cdot R_{pg_j}(\underline{0})\}.$$

Despite the complicated form taken by the flux covariance tensor above, it appears that it should remain invariant under the transformations which leave $p(\underline{x})$ and $g_i(\underline{x})$ jointly invariant. Note that $p(\underline{x})$ is just the scaled perturbation of the conductivity $K(\underline{x})$. Without any rigorous proof, we will assume that $K(\underline{x})$ and $g_i(\underline{x})$ are jointly axisymmetric random fields (we know this is true for the log-conductivity and for $g_i(\underline{x})$ separately). Accordingly, the flux covariance should remain invariant under rotations-reflexions in the transverse plane, and reflexions through the longitudinal direction as explained earlier. This also implies that $R_{q_i q_j}$ is a symmetric tensor. For completeness, observe that $R_{q_i q_j}$ given by (4.19) does satisfy the mass conservation relation (4.8) as required.

Now, by using only the symmetry properties due to statistical isotropy (invariance to rotations and reflexions, as defined above) as well as the mass conservation relation, it is possible to come up with the general form of the covariance or spectrum of $q_i(\underline{x})$ independently of the detailed formula given in (4.19). For instance, Batchelor (1953) gives the general form of the covariance tensor of a vector field which is *statistically axisymmetric*. By applying his results [Eqs. 3.3.9, page 43] and adding the condition of invariance under reflexions through the axis of symmetry, we obtain:

$$R_{q_i q_j}(\xi) = A(\xi_1, r)\,\frac{\xi_i \xi_j}{r^2} + B(\xi_1, r)\,\delta_{ij} + C(\xi_1, r)\,\delta_{1i}\delta_{1j} + D(\xi_1, r)\cdot(\xi_i \delta_{1j} + \xi_j\,\delta_{1i}) \tag{4.20}$$

where r is the radial lag distance $(\xi_2^2 + \xi_3^2)^{1/2}$ in the transverse plane, and A,B,C,D are even functions of ξ_1. Note that the case of spherical isotropy for $R_{q_i q_j}(\xi)$ obtains by taking $C = D = 0$. This suggest that the A and B terms account for the fully isotropic part of velocity fluctuations, while the C and D terms account for those fluctuations driven by the mean head gradient (the driving force responsible for anisotropic behavior).

Similarly, by using properties of the Fourier Transform, it can be shown that the tensor of spectral densities $S_{q_i q_j}(\underline{k})$ must be of the same form as (4.20), namely:

$$S_{q_i q_j}(\underline{k}) = A \cdot \frac{k_i k_j}{k^2} + B\cdot\delta_{ij} + C \cdot \delta_{1i}\delta_{1j} + D\cdot(k_i\delta_{1j} + k_j\delta_{1i}) \tag{4.21}$$

where A, B, C, D are functions of k_1 and $k_R = (k_2^2 + k_3^2)^{1/2}$, and are even in k_1.

We now apply the mass conservation condition in its spectral form (4.7) to obtain additional conditions on A, B, C and $D(k_1, k_R)$ as follows. Combining equation (4.7):

$$k_i \, S_{q_i q_j}(\underline{k}) = 0$$

with equation (4.21) above gives:

$$\begin{aligned} &(A + B)\, k_j \\ &+ (k^2 \cdot D + k_1 \cdot C)\, \delta_{1j} \\ &+ k_1\, k_j\, D = 0 \end{aligned}$$

where $k^2 = k_1^2 + k_2^2 + k_3^2 = k_1^2 + k_R^2$, and $j = 1,2,3$ respectively. These conservation conditions on A,B,C,D can be rewritten as follows:

$$\begin{aligned} &(1): \quad k_1 \cdot A + k_1 B + k_1 C + (k^2 + k_1^2) D = 0 \\ &(2): \quad k_2 \cdot A + k_2 B \qquad + k_1 k_2 D \quad = 0 \qquad (4.22) \\ &(3): \quad k_3 \cdot A + k_3 B \qquad + k_1 k_3 D \quad = 0 \end{aligned}$$

Multiplying each equation by k_i and summing, we obtain the equivalent system of equations:

$$A + B + \frac{k_1^2}{k^2} C + 2k_1 D \qquad = 0$$

$$A + B + C + \frac{k^2+k_1^2}{k_1^2} k_1 D = 0$$

$$A + B + k_1 D \qquad = 0.$$

This leads after some manipulations to just two independent relations:

$$k^2 D + k_1 C = 0$$

$$A + B = \frac{k_1^2}{k^2} C. \tag{4.23}$$

Plugging (4.23) into (4.21) finally gives the general form of the flux spectrum $S_{q_i q_j}$ that satisfies the properties of axial symmetry (as defined earlier) and mass conservation:

$$\boxed{\begin{aligned} S_{q_i q_j}(\underline{k}) = &\left[\frac{k_i k_j}{k^2} - \delta_{ij}\right] \cdot A(k_1, k_R) \\ &+ \left[\delta_{1i}\delta_{1j} + \frac{(k_1)^2}{k^2}\delta_{ij} - \frac{k_1 k_i}{k^2}\delta_{1j} - \frac{k_1 k_j}{k^2}\delta_{1i}\right] \cdot C(k_1, k_R) \end{aligned}} \tag{4.24}$$

where k is the spherical-radial wavenumber, and k_R is the cylindrical-radial wavenumber in the transverse plane (k_2,k_3). Note again that A and C are even functions of k_1.

The general result given in (4.24) may now be compared to specific solutions, such as the first order spectral solutions derived in Chapter 3. Equation (4.24) also applies to isotropic turbulence by taking $C = 0$ (recall that the case $C = 0$ corresponds to spherical symmetry). In this case, equation (4.24) gives the correct result, with A(k) being the 3D radial spectrum of kinetic energy (see for instance Monin and Yaglom, 1965, Eq. 12.73). Let us now compare equation (4.24) with the first order spectral solution (3.18) for the case where the spectrum of log-conductivities is isotropic ($S_{ff} = S_{ff}(k)$). Let us first decompose both equations (4.24) and (3.18) for a term by term comparison. This is shown below, denoting $S_{q_iq_j}$ the general solution and $S^{(1)}_{q_iq_j}$ the first order solution:

$$S_{q_iq_j}(\underline{k}) =$$

$$C\left\{\delta_{i1}\delta_{j1} - \delta_{i1}\frac{k_1k_j}{k^2} - \delta_{j1}\frac{k_1k_i}{k^2} + \frac{A}{C}\cdot\frac{k_ik_j}{k^2} + \left[\frac{(k_1)^2}{k^2} - \frac{A}{C}\right]\delta_{ij}\right\} \qquad (4.25)$$

$$S^{(1)}_{q_i q_j}(\underline{k}) =$$

$$K_G^2 \, J_i^2 \, S_{ff}(k) \cdot \left\{ \delta_{i1}\delta_{j1} - \delta_{i1} \frac{k_1 k_j}{k^2} - \delta_{j1} \frac{k_1 k_i}{k^2} + \frac{(k_1)^2}{k^2} \cdot \frac{k_i k_j}{k^2} \right\}. \tag{4.26}$$

The comparison shows that the first order solution $S^{(1)}_{q_i q_j}$ satisfies the general form of the spectrum given in (4.24) or (4.25), with the choice:

$$\begin{aligned} C(k_1, k_R) = C(k) \quad &= K_G^2 \, J_i^2 \, S_{ff}(k) \\ A(k_1, k_R) = A(k_1, k) &= \frac{(k_1)^2}{k^2} \cdot C(k). \end{aligned} \tag{4.27}$$

The form of the flux covariance tensor can also be found by applying a Fourier Transform to both sides of (4.24). It is worth noting that the flux covariance $R_{q_i q_j}(\underline{\xi})$ as well as the head gradient covariance $R_{g_i g_j}(\underline{\xi})$ must vanish along certain lines or planes when $i \neq j$ (*see Figure (4.1) above*). This is due solely to certain symmetries under rotations or reflexions, and the same should hold for any solution satisfying the stationarity hypothesis. Again, we find that the first order solutions of Chapter 3 do satisfy the relations depicted in Figure (4.1).

In summary, we have obtained the general form of the spectrum of the flux vector independently of any particular perturbation approximation, assuming only that the log-conductivity field is statistically isotropic, and the flow field is statistically homogeneous (stationary). This general solution includes as a special case the first order spectral solutions of Chapter 3 (Bakr et al. 1978, Gelhar and Axness 1983). More generally, we believe that any approximate solution should yield a flux spectrum of the form (4.24) in order to be consistent with the basic statistical properties of the governing flow equation.

Finally, it is worth noting that most of the results in this section may be applied to the 2D *isotropic case* as well, by letting $i = (1,2)$ rather than $i = 1,2,3$ in tensorial expressions. For example, applying (4.24) with $i = (1,2)$ gives the general form of the $S_{q_i q_j}$ spectrum in the 2D case, with A and C even functions of $\underline{k} = (k_1, k_2)$. In fact, we will show that the form of the solutions in the special 2D case can be narrowed down further by using the special symmetry inherent to the 2D space -- leading to a conjugacy relation between flux and head gradient. This is examined next.

4.2.4 Conjugacy property for 2D flow in isotropic media:

For the case of flow in a two-dimensional random porous medium, we use the streamfunction formulation to show that the stochastic flow field must satisfy a conjugacy condition (in probability) between the flux and head gradient vectors.

For details on the streamfunction formulation, the reader is referred to Bear (1972), among others. Briefly, the streamlines are defined as the set of curves tangent to the velocity field at every point in space. Note that the velocity and flux vectors are equivalent in a medium of constant porosity (normalized to unity for convenience). Thus the equation for the streamlines can be written in terms of the flux vector field as follows:

$$\frac{dx_1}{Q_1(\underline{x})} = \frac{dx_2}{Q_2(\underline{x})}.$$

The streamfunction $\Psi(\underline{x})$ is defined as:

$$Q_1(\underline{x}) = -\frac{\partial \Psi}{\partial x_2}$$

$$Q_2(\underline{x}) = +\frac{\partial \Psi}{\partial x_1}. \qquad (4.28)$$

It is easily seen that the curves $\Psi(\underline{x}) = c$ describe the set of streamlines in the flow. The equipotentials $H(\underline{x}) = c$ similarly describe the level curves of the hydraulic head ("potential") field. Finally, note that the flux vector in (4.28) automatically satisfies the conservation equation $\partial Q_i/\partial x_i = 0$.

We now seek a flow equation based solely on the streamfunction Ψ. The Darcy equation implies that $(\underline{Q}/K)$ is a potential vector:

$$\underline{Q}/K = - \underline{\nabla} H$$

whose curl must vanish, i.e.:

$$\underline{\nabla} \times (\underline{Q}/K) = \underline{0}.$$

In fact, only the third component of the curl is of interest here, giving:

$$\frac{\partial}{\partial x_1} (Q_2/K) - \frac{\partial}{\partial x_2} (Q_1/K) = 0. \qquad (4.29)$$

Plugging (4.28) in (4.29) leads to the required Ψ-based equation:

$$\underline{\nabla} \left(\frac{1}{K} \underline{\nabla} \Psi\right) = 0. \qquad (4.30)$$

Equivalently, by using the log-conductivity $F(\underline{x}) = \ell n\, K(\underline{x})$, we obtain the desired equation for the streamfunction:

$$\nabla^2\Psi - \underline{\nabla}F \cdot \underline{\nabla}\Psi = 0. \qquad (4.31)$$

In comparison, the equivalent equation for the hydraulic head "potential" was:

$$\nabla^2 H + \underline{\nabla}F \cdot \underline{\nabla}H = 0. \qquad (4.32)$$

The idea of conjugacy arises from the observation that the governing equation for the streamfunction Ψ can be obtained simply by reversing the sign of $F(\underline{x}) = \ell n\, K(\underline{x})$ in the equation governing the head field. We now examine the implications of this "duality" in the case where the log-conductivity is a Gaussian isotropic random field in 2D space.

Consider the case of a finite square domain, with fixed heads on two opposite boundaries, and zero normal flux on the other boundaries (Figure 4.2). We now use equation (4.28), along with the Darcy equation, to express these boundary conditions in terms of the streamfunction. This is summarized below:

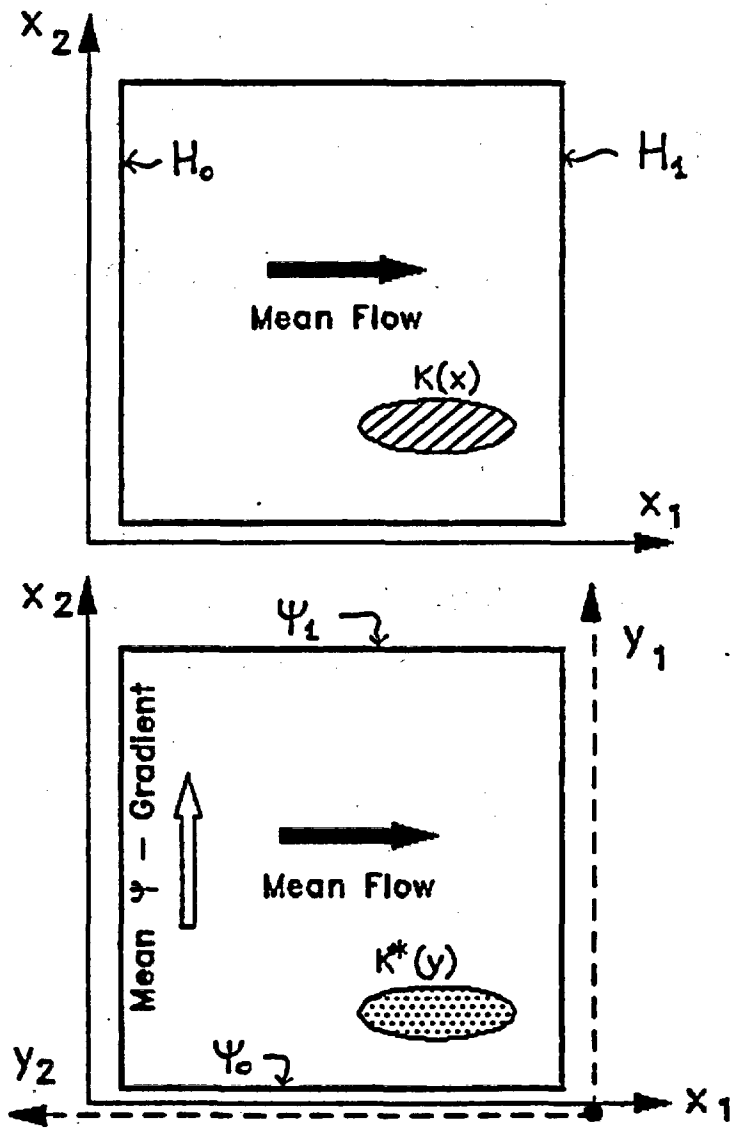


Figure 4.2 Illustration of the conjugacy property for stochastic flow in a 2D isotropic medium (K^* is the dual conductivity with respect to K)

(i) On the fixed head boundaries (e.g., at $x_1 = 0$):

$$H(0,x_2) = H_o \Rightarrow \frac{\partial H}{\partial x_2}(0,x_2) = 0 \Rightarrow \frac{\partial \Psi}{\partial x_2}(0,x_2) = 0.$$

(ii) On the zero-flux boundaries (e.g. at $x_2 = 0$):

$$Q_2(0,x_1) = 0 \Rightarrow \frac{\partial \Psi}{\partial x_1}(0,x_1) = 0 \Rightarrow \Psi(0,x_1) = \Psi_o.$$

This clearly shows that the boundary conditions for the streamfunction equation are of the same type as those for the head equation, provided a 90 degree rotation of the flow domain (see Figure 4.2). Furthermore, observe that the global hydraulic gradient is, by construction:

$$\frac{-(H_1 - H_o)}{L} = \bar{J}_1$$

while the global streamfunction gradient is:

$$\frac{-(\Psi_1 - \Psi_o)}{L} = \frac{1}{L}\int_{x_2=0}^{x_2=L} Q_1 \, dx_2 = \bar{Q}_1.$$

Based on these remarks, the boundary value problem for 2D flow can be expressed indifferently in terms of H or Ψ as shown below. First, we need to define new dimensionless variables:

$$\tilde{H} = \frac{H-H_o}{\bar{J}_1 L}$$

$$\tilde{\Psi} = \frac{\psi-\psi_o}{\bar{Q}_1 L}. \tag{4.33}$$

Second, let us define the *dual conductivity*, dual log-conductivity, and dual log-conductivity perturbation as follows:

$$K^*(x_1,x_2) = 1/K(-x_2,+x_1)$$

$$F^*(x_1,x_2) = -F(-x_2,+x_1) \tag{4.34}$$

$$f^*(x_1,x_2) = - f(-x_2,+x_1).$$

It is easily seen that equations (4.31) and (4.32) are mutually "conjugate", or "dual", i.e., the normalized streamfunction $\tilde{\Psi}$ is solution of the conjugate boundary-value problem involving the dual conductivity field as shown below:

$$\frac{\partial^2\tilde{\Psi}}{\partial y_1 \partial y_1} + \frac{\partial F^*}{\partial y_1} \cdot \frac{\partial \tilde{\Psi}}{\partial y_1} = 0$$

$$\tilde{\Psi}(0,y_2) = 0; \quad \tilde{\Psi}(L,y_2) = 1 \tag{4.35}$$

$$\frac{\partial \tilde{\Psi}}{\partial y_2}(y_1,0) = \frac{\partial \tilde{\Psi}}{\partial y_2}(y_1,L) = 0$$

where $(y_1, y_2) = (x_2, -x_1)$. The equation for the normalized head potential $\tilde{H}$ is identical:

$$\frac{\partial^2 \tilde{H}}{\partial x_1 \partial x_1} + \frac{\partial F}{\partial x_1} \frac{\partial \tilde{H}}{\partial x_1} = 0$$

$$\tilde{H}(0, x_2) = 0; \quad \tilde{H}(L, x_2) = 1 \qquad (4.36)$$

$$\frac{\partial \tilde{H}}{\partial x_2}(x_1, 0) = \frac{\partial \tilde{H}}{\partial x_2}(x_1, L) = 0.$$

The next step in this analysis consists in letting the domain size L become infinite, while the mean hydraulic gradient $\bar{J}_1$ remains finite. Furthermore, we now use the assumption that $F(\underline{x})$ is a stationary, Gaussian, and isotropic random field. Taken together, statistical isotropy and normality imply that the log-conductivity perturbation (f) must have all its moments invariant under rotations and reflexions, as well as invariant under the transformation $f \to -f$ (due to the symmetry of the Gaussian distribution). This implies that the dual field $f^*(\underline{x})$ defined in (4.34) is identical in probability to $f(\underline{x})$. Thus, equations (4.35) and (4.36) imply that the normalized random fields $\tilde{\Psi}(\underline{y})$ and $\tilde{H}(\underline{x})$, expressed in different coordinate systems, are identical in probability. Note that the $\underline{y}$ coordinate system obtains by rotating $\underline{x}$, such that $(y_1, y_2) = (-x_2, +x_1)$.

The above results can be simply stated in one single equation:

$$\boxed{\tilde{\Psi}(-x_2,+x_1) \underset{=}{\Delta} \tilde{H}(x_1,x_2)} \qquad (4.37)$$

where the "$\underset{=}{\Delta}$" sign stands for equality in probability, meaning that all the moments up to arbitrary order must be equal. The above equation simply means that the random patterns of streamlines and equipotentials are statistically identical in a 2D isotropic medium, provided proper normalization of variables and rotation of the coordinates by a 90 degree angle. This is the property we call "conjugacy", to be understood in a statistical sense. Thus, according to equation (4.37), the streamfunction and head potential are statistically "conjugates" of each other.

We now use the conjugacy property (4.37) along with the flux-streamfunction relation (4.28) to show that the flux and head gradient vectors must also be conjugate. First, by plugging (4.33) in (4.37) we obtain, in terms of dimensional quantities:

$$\Psi(y_1,y_2) - \Psi_o \underset{=}{\Delta} \bar{Q}_1 \cdot \frac{H(x_1,x_2)-H_o}{\bar{J}_1}$$

where $(y_1, y_2) = (-x_2, +x_1)$. Using (4.28) and applying standard differentiation rules, this gives:

$$Q_i(x_1, x_2) \triangleq - \frac{\bar{Q}_i}{\bar{J}_i} \cdot \frac{\partial H}{\partial x_i} (x_2, -x_1) \tag{4.38}$$

where the gradient vector $\partial H/\partial x_i$ should be evaluated by differentiating H with respect to the (x_1, x_2) system before substituting $(x_2, -x_1)$. Equation (4.38) shows that the flux and head gradient vectors, normalized by their mean values, are statistically identical upon rotation of the coordinate system by a 90 degree angle.

We now use this important result to obtain an exact relation between the covariance functions or the spectral density functions of the flux and head gradient vector fields. Equation (4.38) immediately leads to:

$$\begin{aligned} R_{q_i q_j}(\xi_1, \xi_2) &= \hat{K}^2 \cdot R_{g_i g_j}(\xi_2, -\xi_1) \\ S_{q_i q_j}(k_1, k_2) &= \hat{K}^2 \cdot S_{g_i g_j}(k_2, -k_1) \end{aligned} \tag{4.39}$$

where $\hat{K} = \bar{Q}_1/\bar{J}_1$ is the *effective conductivity*. Again note that the tensors $R_{g_i g_j}$ on the right-hand-side should be evaluated in the (ξ_1,ξ_2) system before substituting for $(\xi_2,-\xi_1)$: similarly for $S_{g_i g_j}$.

Finally, the head spectrum can be introduced on the right hand side of equation (4.39). Indeed, using the fact that $g_i(\underline{x})$ is a potential vector (gradient of the scalar field $h(\underline{x})$), one may express directly the flux spectrum in terms of the head spectrum as shown below:

$$S_{q_i q_j}(\underline{k}) = \hat{K}^2 \cdot k_i' k_j' \, S_{hh}(\underline{k}')$$

where $(k_1', k_2') = (k_2,-k_1)$. To obtain a more explicit relation, let us rewrite the transformed wavenumber system as:

$$k_i' = -(1-\delta_{i1})\, k_1 + (1-\delta_{i2})\, k_2.$$

This gives:

$$S_{q_i q_j}(k_1,k_2) =$$

$$\hat{K}^2 \cdot \{(1-\delta_{i1})(1-\delta_{j1})\, k_1^2 + (1-\delta_{i2})(1-\delta_{j2})\, k_2^2$$
$$- [(1-\delta_{i1})\,(1-\delta_{j2}) + (1-\delta_{i2})\,(1-\delta_{j1})]k_1 k_2\} \cdot S_{hh}(k_2,-k_1).$$

This expression can be simplified further by defining the radial wavenumber $k = (k_1^2+k_2^2)^{1/2}$ and using certain properties of the Kronecker symbol such as:

$$(1-\delta_{i1})(1-\delta_{j2}) + (1-\delta_{i2})(1-\delta_{j1}) = 1-\delta_{ij}.$$

This gives finally a general expression for the flux spectrum in a 2D isotropic medium, independently of any perturbative approximation other than the stationarity hypothesis:

$$\begin{aligned} S_{q_iq_j}(k_1, k_2) = \hat{K}^2 \cdot \{&(1-\delta_{i1})(1-\delta_{j1})k_1^2 \\ &+ (1-\delta_{i2})(1-\delta_{j2})k_2^2 \\ &- (1-\delta_{ij})\, k_1k_2\} \cdot S_{hh}(k_2, -k_1). \end{aligned} \tag{4.40}$$

Using the radial wavenumber $k = (k_1^2+k_2^2)^{1/2}$, let us give explicitly each component of the symmetric $S_{q_iq_j}$ tensor:

$$\boxed{\begin{aligned} S_{q_1q_1}(k_1, k_2) &= \hat{K}^2 \cdot \left(1- \frac{k_1^2}{k^2}\right) k^2\, S_{hh}(k_2, -k_1) \\ S_{q_2q_2}(k_1, k_2) &= \hat{K}^2 \cdot \left(1- \frac{k_2^2}{k^2}\right) k^2\, S_{hh}(k_2, -k_1) \\ S_{q_1q_2}(k_1, k_2) &= \hat{K}^2 \cdot \left(- \frac{k_1k_2}{k^2}\right) k^2\, S_{hh}(k_2, -k_1) \end{aligned}} \tag{4.40'}$$

Equation (4.40) is an important new result, since it gives the relation between the flux and head spectrum based

solely on the stationarity hypothesis (no small parameter expansion involved). Furthermore, this result can be checked by comparing it to the more general form (4.24) which was derived for 3D isotropic media — and remains valid for 2D isotropic media as well. Indeed the spectrum (4.40) does satisfy the condition (4.24), with $A = 0$ and $C = \hat{K}^2 k^2 S_{hh}(k_2,-k_1)$. Finally, it is instructive to compare (4.40) with the first order spectral solutions from Chapter 3. By applying equations (3.18) for the 2D isotropic case, it is easily seen that the first order approximation for $S_{q_i q_j}$ is just equation (4.40) with $\hat{K}$ replaced by K_G.

We conclude that the first order spectral solutions (3.18) are consistent with the conjugacy condition (4.40), at least up to a constant factor $(\hat{K}/K_G)$. We will show next that the effective conductivity $\hat{K}$ must be precisely equal to the geometric mean K_G in a 2D isotropic medium, assuming only the existence of $\hat{K}$ (i.e., stationary flow field). Therefore, we may conclude here that the first order spectral solution is consistent, in the sense that it satisfies exactly the conjugacy property of two-dimensional isotropic flow systems. Note that any higher-order stationary solutions should also satisfy the conjugacy property in order to be consistent. The conjugacy relations were given by (4.40) for second order moments only, but equation (4.38) could be used to derive similar conjugacy

conditions on higher order moments.

4.2.5 Geometric mean effective conductivity for 2D isotropic media:

In order to complete the previous analysis, we show here that the effective conductivity, if it exists, must be exactly equal to the geometric mean in a 2D isotropic medium. Consider again equations (4.28) to (4.37), and denote Φ the head potential H, and ϕ^* the streamfunction Ψ expressed in the rotated coordinates $(y_1, y_2) = (x_2, -x_1)$. Restating previous results, we have that $\Phi(x)$ and $\phi^*(y)$ are governed by the dual equations:

$$\nabla_x^2 \phi + \underline{\nabla}_x F \cdot \underline{\nabla}_x \phi = 0$$

$$\nabla_y^2 \phi^* + \underline{\nabla}_y F^* \cdot \underline{\nabla}_y \phi^* = 0 \ .$$

Furthermore, the Darcy law:

$$\underline{Q}(\underline{x}) = -K(\underline{x}) \cdot \underline{\nabla} \Phi(\underline{x})$$

has also its dual counterpart:

$$\underline{Q}^*(y) = - K(\underline{y}^*) \cdot \underline{\nabla}_y \phi^*(\underline{y})$$

where, according to previous definitions:

$$K^*(\underline{y}) = - 1/K(\underline{x}). \tag{4.41.a}$$

This leads us to define two "effective conductivities", one for the original ϕ-equation, and one for the dual ϕ^*-equation:

$$\hat{K} = \frac{\langle K(\underline{x}) \cdot \frac{\partial \Phi}{\partial x_1}(x) \rangle}{\langle \frac{\partial \phi}{\partial x_1}(\underline{x}) \rangle}$$

$$\hat{K}^* = \frac{\langle \frac{1}{K^*(\underline{y})} \cdot \frac{\partial \phi^*}{\partial y_1}(\underline{y}) \rangle}{\langle \frac{\partial \phi^*}{\partial y_1}(\underline{y}) \rangle} . \tag{4.41.a}$$

We now borrow an argument from Matheron's indications of a similar proof (Matheron, 1967 and 1984). First, observe that the effective conductivity and its dual satisfy by construction:

$$\hat{K}^* = 1/\hat{K}$$

as can be seen from equations (4.28) to (4.38). Second, note that the effective conductivities $\hat{K}$ and $\hat{K}^*$ must take the form of a functional $\mathcal{F}$ involving possibly all the moments of $K(\underline{x})$. Furthermore, this functional must be of the same form in the two cases, because the governing equations for ϕ and ϕ^* are formally

identical, that is $\phi(x)$ depends on $K(x)$ in the same manner that $\phi^*(y)$ depends on $K^*(\underline{x})$. Whence:

$$\hat{K} = \mathcal{F}(K(\underline{x}))$$

$$\hat{K}^* = \mathcal{F}(1/K(\underline{x})) = 1/\hat{K}. \tag{4.42}$$

The functional in (4.42) must behave linearly with respect to multiplication by a constant, e.g. $\mathcal{F}(\alpha K(\underline{x})) = \alpha\, \mathcal{F}(K(\underline{x}))$: try equation (4.42) with $\alpha K(\underline{x})$. We now use a special property of the log-normal distribution, namely that $K/\langle K\rangle$ and $K^{-1}/\langle K^{-1}\rangle$ have identical distributions in terms of all N-points moments ($N = 1,2,\cdots$). This will in fact hold for any conductivity field whose logarithm has a symmetric distribution. The required result follows directly from equation (4.42), along with the symmetry of the $\ell n\, K$ distribution, and the invariance of $\hat{K}$ with respect to coordinates:

$$\begin{aligned}
\hat{K}^*/\hat{K} &= \mathcal{F}\,(K^{-1}(y)) \\
&= \mathcal{F}(K(y) \cdot \frac{\langle K^{-1}\rangle}{\langle K\rangle}) \\
&= \frac{\langle K^{-1}\rangle}{\langle K\rangle} \cdot \mathcal{F}(K(y)) \\
&= \frac{\langle K^{-1}\rangle}{\langle K\rangle} \cdot \hat{K}\ .
\end{aligned}$$

From the previous identity $\hat{K}^* = 1/\hat{K}$, this gives immediately:

$$\hat{K} = \sqrt{\langle K \rangle / \langle K^{-1} \rangle} \qquad (4.43)$$

Now, by using also the properties of the log-normal distribution:

$$f(\underline{x}) = \ell n \, (K(\underline{x})/K_G)$$

$$K_G = \exp\langle \ell n \, K \rangle \qquad \text{(geometric mean)}$$

$$\langle K \rangle = K_G \, e^{\sigma^2/2} \qquad \text{(arithmetic mean)}$$

$$\langle K^{-1} \rangle^{-1} = K_G \, e^{-\sigma^2/2} \qquad \text{(harmonic mean)}$$

we obtain finally the announced result:

$$\boxed{\hat{K} = K_G \, .} \qquad (4.43')$$

Incidentally, it is interesting to note that a similar proof was obtained for electric networks, by using the concept of a dual conductivity network (Marchant and Gabillard, 1975).

The fact that the effective conductivity is equal to the geometric mean for a 2D isotropic medium with a symmetric probability distribution of $\ell n \, K$, was mentioned by Matheron (1967), Marchant and Gabillard (1975), and Matheron (1984). The proof given above follows and expands on a review published by Matheron (1984). It seems natural to ask whether an exact closed form relation for the effective conductivity could be obtained in

more than 2 dimensions. An interesting formula for the case of statistically isotropic log-normal conductivity in m-dimensional space was suggested by Matheron (1967, Chapt. VI, p. 132), however with no proof. In terms of log-conductivities, Matheron's conjecture can be re-stated as follows:

$$\ell n\ (\hat{K}) = (1 - \frac{1}{m})\ \ell n\ K_A + \frac{1}{m}\ \ell n\ K_H \qquad (4.44)$$

where K_A is the arithmetic mean $\langle K \rangle$, and K_H the harmonic mean $\langle K^{-1} \rangle^{-1}$. This equation is indeed exact for m = 1 and 2 dimensions, giving respectively the harmonic mean and the geometric mean $K_G = \sqrt{K_A \cdot K_H}$. For 3-dimensions, Matheron's conjecture gives exactly the same result as the first order spectral theory: $\hat{K} = K_G \exp(\sigma_f^2/6)$. It is remarkable that these approximations obtained by two different methods match exactly, even though they might be inaccurate for large values of σ_f. Furthermore, as the number of dimensions goes to infinity, the effective conductivity (4.44) converges to its upper bound, the arithmetic mean K_A. Thus, as the "dimensionality" of flow increases from one to infinity, the effective conductivity increases monotically from its lower bound $K_H = K_G\ e^{-\sigma^2/2}$ to its upper bound $K_A = K_G\ e^{+\sigma^2/2}$. This is indeed an attractive feature of Matheron's formula. Incidentally, the proof that the effective conductivity is bounded by K_H and K_A was given by Matheron (1967) based on energy arguments.

In our view, the question of the adequacy of current estimates of the effective conductivity tensor $\hat{K}_{ij}$ for realistic three-dimensional flow problems remains open. However, anticipating the results of numerical simulations with 3D isotropic media (Chapter 6), we stress the fact that the effective conductivity predicted by the spectral theory (or by Matheron's conjecture) agreed very well with the numerical simulation results for a wide range of log-conductivity variability, up to $\sigma_f = 2.3$. Based on this encouraging result, we now investigate how Matheron's conjecture (4.44) could be generalized to include the case of statistically anisotropic media.

4.2.6 Effective conductivity for general 3D anisotropic media:

The proposed generalization of (4.44) is based on the observation that the parameter m should be interpreted as the number of degrees of freedom of fluid particles, rather than the dimensionality of space. When the log-conductivity is a three-dimensional ellipsoidal random field with anisotropic length scales $(\lambda_1, \lambda_2, \lambda_3)$, it seems reasonable to assume that the degree of freedom of flow will depend solely on these three length scales. Furthermore, the first order spectral results obtained by Gelhar and Axness (1983) for various anisotropy ratios (λ_1/λ_3, etc.) suggest that equation (4.44) could be

generalized in the form:

$$\ell n\ \hat{K}_{ii} = \alpha_{ii}\ \ell n\ K_A + \beta_{ii}\ \ell n\ K_H \qquad (4.45)$$

where the coefficients α_{ii} and β_{ii} must be somehow related to the degrees of freedom available for flow in "parallel" mode (K_A) and flow in "perpendicular" mode (K_H), for each of the three principal axes of anisotropy (i=1,2,3).

We now show how α_{ii} and β_{ii} should depend on the length scales ($\lambda_1,\lambda_2,\lambda_3$) for a few special cases, assuming for convenience that the mean head gradient $\underline{J}$ is aligned with one of the principal axes (x_1,x_2,x_3). In this case, $\hat{K}_{ij}$ is a diagonal tensor. The more general case of arbitrary orientation of $\underline{J}$ will follow by using tensorial transformation rules under rotations, assuming that $\hat{K}_{ij}$ indeed behaves like a second rank symmetric tensor (see Matheron 1967, and Gelhar and Axness, 1983).

Let us focus first on the behavior of $\hat{K}_{11}$ in certain special cases where a close-form result is available. The results given below were deduced in part from the work of Gelhar and Axness (1983, Eqs. 4.52-4.60):

(i) $\underline{\lambda_1 = \lambda_2 = \lambda_3}$:

$$\ell n\, \hat{K}_{11} = \frac{2}{3}\, \ell n\, K_A + \frac{1}{3}\, \ell n\, K_H$$

(ii) $\underline{\lambda_1 = \lambda_2,\ \lambda_3 \to 0}$:

$$\ell n\, \hat{K}_{11} = 1 \cdot \ell n\, K_A + 0 \cdot \ell n\, K_H = \ell n\, K_A$$

(iii) $\underline{\lambda_1 = \lambda_2,\ \lambda_3 \to \infty}$:

$$\ell n\, \hat{K}_{11} = \frac{1}{2}\, \ell n\, K_A + \frac{1}{2}\, \ell n\, K_H = \ell n\, K_G$$

(iv) $\underline{\lambda_1 \neq \lambda_2,\ \lambda_3 \to \infty}$:

$$\ell n\, \hat{K}_{11} = \frac{\lambda_1}{\lambda_1+\lambda_2}\, \ell n\, K_A + \frac{\lambda_2}{\lambda_1+\lambda_2}\, \ell n\, K_H.$$

The first case above corresponds to a fully isotropic medium; the second case corresponds to a horizontally stratified medium with isotropic horizontal slices; the third case represents a vertically "stratified" medium analogous to a bundle of independent vertical columns of circular section; and the fourth case is a generalization of the previous one, with vertical columns of ellipsoidal section.

In addition, the case of horizontal flow perpendicular to vertical strata obtains by taking λ_2 and λ_3 infinite which yields $\hat{K}_{11} = K_H$ as expected. Similarly, for flow parallel to horizontal strata, one obtains $\hat{K}_{11} = K_A$ by taking λ_1 and λ_2 infinite.

The other principal components $\hat{K}_{22}$ and $\hat{K}_{33}$ may be obtained in a similar fashion. For instance in the case ($\lambda_1 \neq \lambda_2$, $\lambda_3 \to \infty$) $\hat{K}_{22}$ obtains from $\hat{K}_{11}$ by inverting (λ_1, λ_2), and $\hat{K}_{33}$ by inverting (λ_1, λ_3). The result, given below, follows from the first order spectral analysis of Gelhar and Axness (1983):

$\underline{\lambda_1 \neq \lambda_2,\ \lambda_3 \to \infty:}$

$$\hat{K}_{11} = \frac{\lambda_1}{\lambda_1+\lambda_2} \ell n\ K_A + \frac{\lambda_2}{\lambda_1+\lambda_2} \ell n\ K_H$$

$$K_{22} = \frac{\lambda_2}{\lambda_1+\lambda_2} \ell n\ K_A + \frac{\lambda_1}{\lambda_1+\lambda_2} \ell n\ K_H$$

$$K_{33} = \ell n\ K_A .$$

Taken together, these results indicate that the α_{ii}, β_{ii} are functions of $(\lambda_1, \lambda_2, \lambda_3)$ that follow a few simple rules, listed below:

(1) $\alpha_{ii}(\underline{\lambda})$ is identical to $\alpha_{jj}(\underline{\lambda}')$ with $\underline{\lambda}'$ obtained by interversion of λ_i, λ_j.

(2) $\alpha_{ii}(\underline{\lambda}) + \beta_{ii}(\underline{\lambda}) = 1.$

(3) $\alpha_{11}(\lambda, \lambda, \lambda) = 2/3$ (4.46)

(4) $\alpha_{11}(\lambda, \lambda, 0) = 1$

(5) $\alpha_{11}(\lambda_1, \lambda_2, \infty) = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

These remarks eventually led us to a more general rule satisfying all the requirements in (4.46):

$$\alpha_{11}(\lambda_1,\lambda_2,\lambda_3) = \frac{\lambda_1}{\lambda_1 + \frac{\lambda_2\lambda_3}{\lambda_2+\lambda_3}} \tag{4.47}$$

with the provision that α_{22} and α_{33} can be deduced from α_{11} according to (4.46.1), and $\beta_{11}=1-\alpha_{11}$ according to (4.46.2).

Finally, by using (4.45) (4.46) and (4.47), we obtain a generalization of Matheron's conjecture, applicable to the general case of 3D anisotropic media:

$$\boxed{\begin{aligned} &\ell n\, \hat{K}_{11} = \alpha_{11}(\underline{\lambda}) \cdot \ell n\, K_A + (1-\alpha_{11}(\underline{\lambda}))\ell n\, K_H \\ &\alpha_{11}(\underline{\lambda}) = \frac{\lambda_1^2}{\lambda_1^2 + \frac{\lambda_1\lambda_2\lambda_3}{(\lambda_1+\lambda_2+\lambda_3)-\lambda_1}}. \end{aligned}} \tag{4.48}$$

Equivalently, by using the relations:

$$K_A = K_G\, e^{+\sigma^2/2}$$
$$K_H = K_G\, e^{-\sigma^2/2}$$

valid in the case of a log-normal conductivity field, equation

(4.48) can be expressed in the simpler form:

$$\boxed{\hat{K}_{ii} = K_G \exp\left\{-\frac{\sigma_f^2}{2}\,(1-2\alpha_{ii}(\underline{\lambda}))\right\}} \qquad (4.49)$$

where α_{ii} was defined in equation (4.48) above.

Furthermore, the case where the mean head gradient $\underline{J}$ is not aligned with the anisotropy axis can be resolved by rotating the coordinate system to coincide with $\underline{J}$ and by applying tensorial transformation rules as explained in Gelhar and Axness, 1983. This yields:

$$\boxed{\hat{K}'_{ij} = a_{im}\, a_{jm}\, \hat{K}_{mm}} \qquad (4.50)$$

where $\hat{K}_{mm}$ is given by (4.48) or (4.49). The matrix $[a_{ij}]$ represents the rotation from $\underline{J}$ to the principal axis x_1, and $\hat{K}'_{ij}$ represents the effective conductivity tensor expressed in the (x_1, x_2, x_3) system of principal axes. Note that when the angle $(\underline{J}, x_1)$ is zero, we obtain $\hat{K}'_{ij} = \delta_{im}\, \delta_{jm}\, \hat{K}_{mm}$ as expected, i.e., the effective conductivity is a diagonal tensor in this case.

It is instructive to see how equations (4.48) to (4.50) apply in specific cases of practical interest. We give below the

three principal components of the effective conductivity tensor, according to (4.49), for the case of vertical-to-horizontal anisotropy:

$$\lambda_1 = \lambda_2 = \lambda$$

$$a = \lambda_3/\lambda$$

$$\hat{K}_{11} = \hat{K}_{22} = K_G \exp\left\{+ \frac{\sigma^2}{2} \cdot \frac{1}{1+2a}\right\} \qquad (4.51)$$

$$\hat{K}_{33} = K_G \exp\left\{- \frac{\sigma^2}{2} \cdot \frac{1-2a}{1+2a}\right\}.$$

In particular, the case of imperfect horizontal stratification (horizontal slices) obtains by taking $a < 1$. The perfectly stratified case corresponds to $a \to 0$. The case $a > 1$ is less typical, corresponding to a formation made of vertical elongated lenses of cylindrical shape (or elongated ellipsoids with circular section).

More complex cases, such as those involving three different length scales, can be worked out directly from (4.49). For example, a typical situation may involve a slight anisotropy in the horizontal plane (1:2), and a significant anisotropy in the vertical plane (1:8). Taking $\lambda_1 = 8$, $\lambda_2 = 4$, $\lambda_3 = 1$ and $K_G = 1$, $\sigma_f = 2.3$, the values 8.7, 5.4, and 0.30 for the effective conductivities $\hat{K}_{11}, \hat{K}_{22}, \hat{K}_{33}$, respectively.

In summary, we have extended a conjecture by Matheron (1967) in order to evaluate the effective conductivity in a three-dimensional anisotropic medium characterized by three length scales $(\lambda_1,\lambda_2,\lambda_3)$. The proposed formula (4.48–4.50) matches the results known to be exact for all cases where these are available: arithmetic mean for "parallel flow", harmonic mean for "perpendicular flow", and geometric mean for two-dimensional isotropic flow. In addition, the conjecture also coincides with the results of the first order spectral theory (Gelhar and Axness) in all the special cases examined, such as the case of anisotropy in the horizontal plane with the vertical length scale much larger than the horizontal. Such a general closed form expression for the effective conductivity was not available before, and could be useful for applications.

4.3 New Closed Form Perturbative Solution for the Flux Spectrum:

In this section, we build on our previous analyses of the flux spectrum to suggest a new linearization of the stochastic flow equation for obtaining first order spectral solutions. Specifically, the results obtained for 2D flow, suggest that the flux spectrum in any number of dimensions could be proportional to $\overline{Q_1^2}$ (see equations 4.24, 4.25, and 4.40), rather than the factor $K_G^2\,\overline{J_1^2}$ given by the standard spectral theory (equations 3.18 or 4.26).

Building on this observation, we will show here that the flux spectrum can be obtained directly from a linear second order PDE governing the flux vector q_1. Solving this equation by the first order perturbation method indeed leads to a factor $\overline{Q_1^2}$ rather than $K_G^2\,\overline{J_1^2}$ in the flux spectrum. This is thought to be a more accurate result because of the "linearity" of the stochastic flux-based equation (compared to the "non-linearity" of the stochastic Darcy equation). The numerical experiments of Chapter 6 will confirm the validity of the new linearization approach. We now proceed to develop the new first order solutions in detail.

The flux-based equation can be obtained by applying linear partial differential operators to the Darcy and continuity equations:

$$\begin{aligned} \underline{Q} &= -K\underline{\nabla}H \\ \underline{\nabla}\cdot\underline{Q} &= 0. \end{aligned} \tag{4.52}$$

For clarity of notation, observe that we use the Nable operator for the gradient ($\underline{\nabla}H$) as well as for the divergence (scalar product $\underline{\nabla}\cdot\underline{Q}$). The next step is based on the observation that the vector $\underline{Q}/K$ is a gradient, so that its curl must vanish:

$$\underline{\nabla} \times (\underline{Q}/K) = \underline{0}. \tag{4.53}$$

In addition, from the mass conservation equation (4.52), the gradient of the divergence of $\underline{Q}$ vanishes:

$$\underline{\nabla}(\underline{\nabla}\cdot\underline{Q}) = \underline{0}. \tag{4.54}$$

Note that equations (4.53) and (4.54) express the fact that $\underline{Q}/K$ is a gradient (Darcy law) and $\underline{Q}$ is a divergence free vector (mass conservation). By using the standard rules of vector field operators (see for instance Gradshtein and Rhyzik, 1980, 10.31), equation (4.54) gives:

$$\underline{\nabla}^2\,\underline{Q} + \underline{\nabla} \times (\underline{\nabla} \times \underline{Q}) = \underline{0}\ . \tag{4.55}$$

The curl term in (4.55) can be decomposed as follows:

$$\begin{aligned}\underline{\nabla} \times (\underline{\nabla} \times \underline{Q}) &= \underline{\nabla} \times (\underline{\nabla} \times (K \cdot \underline{Q}/K)) \\ &= \underline{\nabla} \times [\underline{K}\ \underline{\nabla} \times (\underline{Q}/K) + \underline{\nabla}(K) \times Q/K] \\ &= \underline{\nabla} \times (\tfrac{1}{K}\ \underline{\nabla}K \times \underline{Q})\end{aligned}$$

where the last step obtains by using equation (4.53). Plugging the above identity in (4.55) gives:

$$\underline{\nabla}^2\,\underline{Q} + \underline{\nabla} \times (\tfrac{1}{K}\ \underline{\nabla}\ (K) \times \underline{Q}) = 0\ .$$

This equation is nonlinear in K, but can be made linear in the

log-conductivity perturbation $f(\underline{x}) = \ell n(K(\underline{x})/K_G)$. Indeed, by using $K = K_G\, e^f$ it comes:

$$\boxed{\underline{\nabla}^2 \underline{Q} + \underline{\nabla} \times (\underline{\nabla} f \times \underline{Q}) = \underline{Q}} \,. \qquad (4.56)$$

Equation (4.56) constitutes a system of three scalar second order partial differential equations governing the flux vector components. The most remarkable feature of this simple equation is its linearity with respect to the log-conductivity field $f(\underline{x})$. In comparison, the standard spectral solutions were based on the "nonlinear" Darcy equation, of the form:

$$\underline{Q} = -K_G\, e^{f(\underline{x})} \cdot \underline{\nabla} H$$

where the exponential e^f is obviously a strongly nonlinear function of f.

Equation (4.56) can be made more explicit by decomposing the curl term as follows:

$$\underline{\nabla} \times (\underline{\nabla} f \times \underline{Q}) = \underline{\nabla} f (\underline{\nabla} \cdot \underline{Q}) - \underline{Q}(\underline{\nabla} \cdot \underline{\nabla} f) - (\underline{\nabla} f \cdot \underline{\nabla})\underline{Q} + (\underline{Q} \cdot \underline{\nabla})\underline{\nabla} f .$$

Using again $\underline{\nabla} \cdot \underline{Q} = 0$ and plugging the above expression in (4.56) yields:

$$\boxed{\underline{\nabla}^2\underline{Q} - (\underline{\nabla}f \cdot \underline{\nabla})\underline{Q} = \underline{Q}\nabla^2 f - (\underline{Q} \cdot \underline{\nabla})\underline{\nabla}f} \qquad (4.57)$$

where $\underline{\nabla}^2$ is the vector-Laplacian defined earlier, and ∇^2 is the usual scalar Laplacian. For convenience, let us express (4.57) in tensorial form:

$$\boxed{\frac{\partial^2 Q_i}{\partial x_j \partial x_j} - \frac{\partial f}{\partial x_j}\frac{\partial Q_i}{\partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_j} Q_i - \frac{\partial^2 f}{\partial x_i \partial x_j} Q_j.} \qquad (4.58)$$

It is easy to verify that (4.57) or (4.58) is indeed a valid governing equation for flow, by plugging $\underline{Q} = -K\underline{\nabla}H$ and using $\underline{\nabla} \cdot \underline{Q} = 0$ for mass conservation.

The linear form of the flux-based equation (4.58) suggests that the standard spectral solutions for the flux vector field could be improved by using (4.58) rather than the Darcy equation for a perturbative spectral analysis of the flux vector. In keeping with the basic premises of the spectral method, we assume now that the input log-conductivity as well as the output flux vector are stationary random fields. Defining the flux perturbation:

$$q_i = Q_i - \bar{Q}_i$$

and using the stationarity assumption ($\bar{Q}_i$ constant), we obtain the mean equation by averaging (4.58):

$$- \left\langle \frac{\partial f}{\partial x_j} \frac{\partial q_i}{\partial x_j} \right\rangle = \left\langle \frac{\partial^2 f}{\partial x_j \partial x_j} q_i \right\rangle - \left\langle \frac{\partial^2 f}{\partial x_i \partial x_j} q_j \right\rangle . \qquad (4.59)$$

Note that the mean flux components $\bar{Q}_i$ cancel out due to the fact that $\langle f \rangle = 0$ and $\partial \bar{Q}/\partial x = 0$. Next, the equation for the flux perturbation obtains by subtracting (4.59) from (4.58). This gives after some manipulations:

$$\frac{\partial^2 q_i}{\partial x_j \partial x_j} - \frac{\partial^2 f}{\partial x_j \partial x_j} \cdot \bar{Q}_i + \frac{\partial^2 f}{\partial x_i \partial x_j} \cdot \bar{Q}_j = 0(\sigma_f^2). \qquad (4.60)$$

The term that was neglected on the right hand side of (4.60) is the second-order perturbation:

$$\mathcal{P}\left\{ \frac{\partial f}{\partial x_j} \frac{\partial q_i}{\partial x_j} + \frac{\partial^2 f}{\partial x_j \partial x_j} \cdot q_i - \frac{\partial^2 f}{\partial x_i \partial x_j} \cdot q_j \right\} \qquad (4.61)$$

where we used the operator $\mathcal{P}$ to denote the perturbation of a random quantity: $\mathcal{P}(Y) = Y - \langle Y \rangle$.

Now, the first order spectral solution for the flux field is readily obtained by applying the spectral representation theorem as in Chapter 3. This gives, from (4.60) with the terms on the right-hand side neglected:

$$\boxed{dZq_i(\underline{k}) = \{\bar{Q}_i - \frac{k_i(k_m\bar{Q}_m)}{(k_m k_m)}\}\, dZ_f(\underline{k})}$$

$$\boxed{S_{q_i q_j}(\underline{k}) = \{\bar{Q}_i - \frac{k_i(k_m\bar{Q}_m)}{k^2}\} \cdot \{\bar{Q}_j - \frac{k_j(k_m\bar{Q}_m)}{k^2}\}\, S_{ff}(\underline{k}).} \qquad (4.62)$$

Note that the flux spectrum in (4.62) is entirely determined once the effective conductivity tensor is known. Recall that the effective conductivity relates the mean flux to the mean head gradient via:

$$\bar{Q}_i = \hat{K}_{im}\, \bar{J}_m.$$

For the particular case where the mean head gradient coincides with the principal axis of the $\hat{K}_{ij}$ tensor (say x_1), equation (4.62) simplifies to:

$$\boxed{S_{q_i q_j}(\underline{k}) = (\bar{Q}_1)^2 \cdot \left\{\delta_{1i} - \frac{k_1 k_i}{k^2}\right\} \cdot \left\{\delta_{1j}\, \frac{-k_1 k_j}{k^2}\right\} S_{ff}(\underline{k})} \qquad (4.63)$$

where $\bar{Q}_1 = \hat{K}_{11}\,\bar{J}_1$ should be used.

Comparing now (4.63) to the result from the standard spectral result (3.18), it appears that equation (4.63) obtains from (3.18) by replacing the factor $(K_G J_1)$ by $\bar{Q}_1$. Thus, the shape of the spectrum is not affected by the new approach, and the flux correlation functions will remain unchanged. On the other hand, the flux standard deviations σ_{q_1} previously obtained in Chapter 3 must now be multiplied by the factor $(\hat{K}_{11}/K_G)$ according to the new approach.

It is instructive to compare (4.63) to the standard spectral theory (3.18) by examining the behavior of the ratio $\hat{K}_{11}/K_G$. First note that both (3.18) and (4.63) give $S_{qq} = 0$ for the pathological one-dimensional case (q must be constant in order to satisfy mass conservation in one dimension). For the 2D isotropic case, (3.18) and (4.63) coincide exactly since $\hat{K}_{11} = K_G$ in this case. For the 3D isotropic case we have $\hat{K}_{11}/K_G = \exp(\sigma_f^2/6)$ and the discrepancy will increase with σ_f. The discrepancy between (3.18) and (4.63) will be even higher in the case of strongly anisotropic media, such as flow perpendicular to perfect stratification $(\hat{K}_{11}/K_G = e^{-\sigma^2/2})$ and flow parallel to perfect stratification $(\hat{K}_{11}/K_G = e^{+\sigma^2/2})$.

Let us illustrate these remarks for two simple cases using the notation $\sigma_q^{(1)}$ for the standard theory (3.18) and $\sigma_q^{(2)}$ for the new result (4.63). For a 3D isotropic medium with σ_f=1, the ratio of $\sigma_q^{(2)}/\sigma_q^{(1)}$ is about 1.2. In the case of flow parallel to stratification in a perfectly stratified medium, this ratio would rise to about 1.7. We conclude that the discrepancy between (3.18) and (4.63) is quite significant as far as the flux standard deviations are concerned. It seems reasonable to assume that the most accurate formula is (4.63), since it is based on a "linear" equation governing the flux vector. The forthcoming numerical experiments (Chapter 6) will confirm that the standard solution (3.18) appears to underestimate the flux variances, whereas (4.63) is in better agreement with numerical results.

Incidentally, the new result obtained here indicates that the spectral solutions of stochastic solute transport should be modified as well. In particular, the longitudinal macrodispersivity for 3D solute transport given by Gelhar and Axness (1983):

$$A_{11} = \frac{\sigma_f^2 \lambda}{\gamma^2} \quad (\gamma = \hat{K}_{11}/K_G) \tag{4.64}$$

should now be revised according to equation (4.63), as follows:

$$A_{11} = \sigma_f^2 \lambda. \tag{4.65}$$

It is interesting to note that the new macrodispersivity given by (4.65) increases monotonically with σ_f, whereas the expression (4.64) of Gelhar and Axness presents a maximum at some positive value of σ_f. Beyond this value, the macrodispersivity would decrease with increasing variability. This behavior could not be explained on physical grounds, which makes the new solution (4.65) more attractive.

Along the same lines, note that the coefficient of variation of the flux component ($\sigma_{q_1}/\bar{Q}_1$) also presents a maximum with respect to σ_f according to the standard solution (3.18). In contrast, this coefficient increases monotonically like σ_f when (4.63) is used instead. Again, there seems to be no intuitive explanation for the occurrence of a maximum in $\sigma_{q_1}/\bar{Q}_1$ as σ_f increases. We conclude that the proposed spectral solution for the flux (4.63) seems preferable. This conclusion is also justified by observing that the linearization of e^f used in the standard spectral perturbation method was not required in the present approach. However, it should be recognized that other linearization approximations involved in the solute transport equation were not eliminated or improved by the present approach.

4.4 Finite Size Effects: Band-Pass Self-Similar Spectra, Spectral Conditioning and Uncertainty.

4.4.1 Motivation and approaches:

One of the major difficulties encountered in the application of spectral methods to field cases arises from the fact that groundwater flow fields (and solute concentration fields) are not, in reality, *stationary* random fields over infinitely large domains. Here, the term "stationary" is understood in the usual sense of statistical homogeneity, or translation invariance in probability. Recall that the spectral solution method, based on the representation theorem of random functions in Fourier space, required assuming the existence of stationary solutions over infinite domains in order to solve the governing stochastic equations in closed form (Chapter 3, and previous sections of Chapter 4). Other methods were developed in the literature to deal explicitly with non-stationary problems, and they may be more appropriate in cases of irreducible non-stationary behavior (e.g., drawdown near a pumping well). For instance the approximate Green's function method of Dagan (1982), and the numerical solution of approximate moment equations (Townley 1983, McLaughlin 1985) can be used to obtain non-stationary solutions in the case of small variability. These solution methods have their advantages. However, they do not possess the analytical simplicity of the infinite domain spectral approach (see literature review in Chapter 2).

Due to the approximations involved, the spectral method seems best adapted to the study of "*large scale*" flow and transport phenomena where the effect of local inhomogeneities on the overall pattern is minimal. However, field contamination studies usually focus on phenomena occurring on a finite scale. The scale in question may well be imposed by the geologic structure itself, or by policy considerations (target time for plume prediction). Moreover, certain physical phenomena like solute transport or unsaturated infiltration from *local sources* necessarily take place over some finite scale *evolving in time*.

In addition, experimental studies published in the literature indicate that the statistical properties of conductivities or transmissivities measured at different scales (or by different researchers!) may vary greatly. This is especially true for the "correlation scale", as noted in the previous data review in Chapter 2. The reader is referred in particular to section 2.3 for more complete references on field observations and data collection.

In our view, these difficulties indicate that, at least in some cases, the apparent correlation structure of the conductivity field inferred from field data depends on the scale of the observation. This suggests that a similar phenomenon may occur physically in transient flow and/or mass transport systems

dominated by local sources. Consider for instance a contamination plume originating from a local source, and convected in a steady groundwater flow field through a heterogeneous porous formation. This is illustrated in Figure 4.3 (top). As the plume grows and invades larger regions of the porous formation, it "responds" to larger and larger heterogeneities. At early times, the plume's spreading process is being "excited" by small heterogeneities which appear large with respect to the plume. Later on, the same small heterogeneities appear as mere fluctuations, with respect to the larger plume. In summary, as the plume grows and spreads, there is a change in the typical size of those heterogeneities that affect the global trends of the plume, and those that contribute to random-like mechanical dispersion within the plume. Accordingly, the apparent macrodispersion of the plume is likely to depend on the size or time scale of interest.

The phenomenon of scale dependence may also play a role for "large scale" characterization of groundwater flow fields. Figure 4.3 (centerpiece) illustrates that the apparent mean and the trend of the log-conductivity may depend on the size of the domain of interest in a given subsurface formation. To illustrate this scale effect, we have used borehole data from the Mont Simon formation (Gelhar, 1976, and Bakr, 1976); the figure shows that, although the mean log-conductivity appears constant

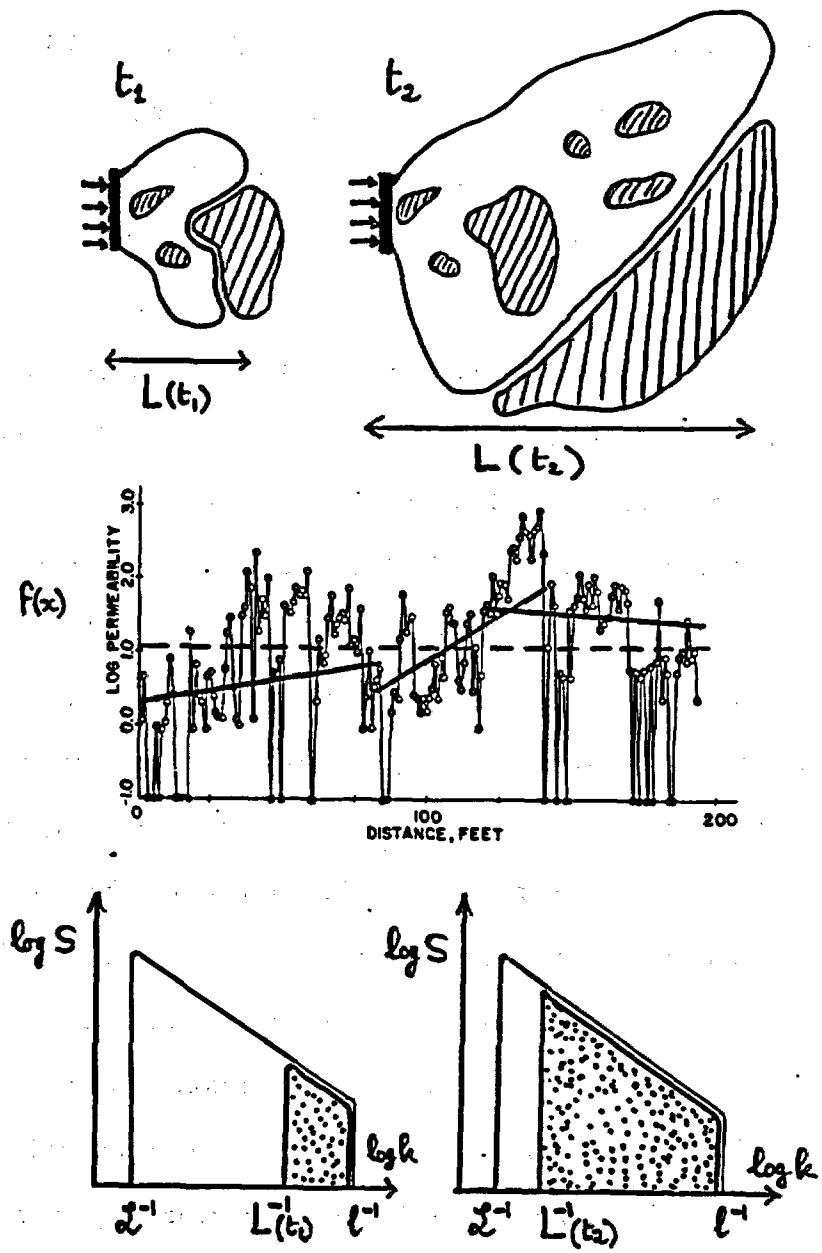


Figure 4.3: Illustration of finite-size effects:
(a) Contamination plume
(b) Log-conductivity field sample function
(c) Band-pass self-similar spectrum

at the scale 200 feet, there exist pronounced trends in the *apparent* mean log-conductivity at the scale 75 feet. This, in turn, suggests that the effective "large scale" conductivity could depend on domain size (and location) within a given regional formation.

We now briefly indicate how the standard spectral theory could be manipulated to incorporate finite-size effects. A first step in this direction consists in using *band-pass* spectra for the log-conductivity field, with a low wavenumber cut-off proportional to the inverse size of the domain ($k_{min} \simeq 1/L$). This is illustrated on the bottom part of Figure 4.3. The high wavenumber cut-off ($k_{max} \simeq 1/\ell$) takes into account the measurement spacing, or possibly the typical scale of conductivity measurements. In addition, the band-pass spectral representation of finite-scale phenomena will be considerably simplified by assuming a *self-similar behavior* of the spectrum within the range of scales of interest. This will be justified shortly by examining some available spectral data. Note, however, that the self-similar behavior may not hold for very large scale phenomena. In the latter case, the infinite-domain spectral theory could be applied safely provided that $L^{-1} \ll \mathcal{L}^{-1}$, where $\mathcal{L}^{-1}$ is the wavenumber below which the spectral content of the log-conductivity becomes negligible. A very large measurement network may be required in order to detect

the actual value of the low wavenumber cut-off $\mathcal{L}^{-1}$, if such a value exists at all.

Finally, the "band-pass" spectral approach could be made more useful if it were possible to incorporate in a simple manner the dependence of the "mean" log-conductivity field with respect to the size of the domain under consideration (see centerpiece in Figure 4.3). This motivated the idea of "spectral conditioning", leading to a clear distinction between *uncertainty* and *spatial variability*. The main features of this new approach will be outlined towards the end of this section, building on the simple band-pass self-similar model.

4.4.2 Band-pass self-similar spectra and field data:

Figures 4.4(a,b,c) display one-dimensional log-conductivity spectra obtained from three sets of borehole data in the Mont Simon aquifer (reproduced from Bakr's thesis, 1976). For Figure 4.4.a, the domain size was $L = 303$ ft, with data spacing $\ell = 1$ ft. The log-spectra density is plotted against the log-frequency in cycles/feet units (wave number = 2π × frequency). The frequency range shown in the figure is approximately $5L^{-1}$ to $0.5\ell^{-1}$ cycles. The straight line superimposed on the spectral data represents a self-similar spectrum with exponent one, that is:

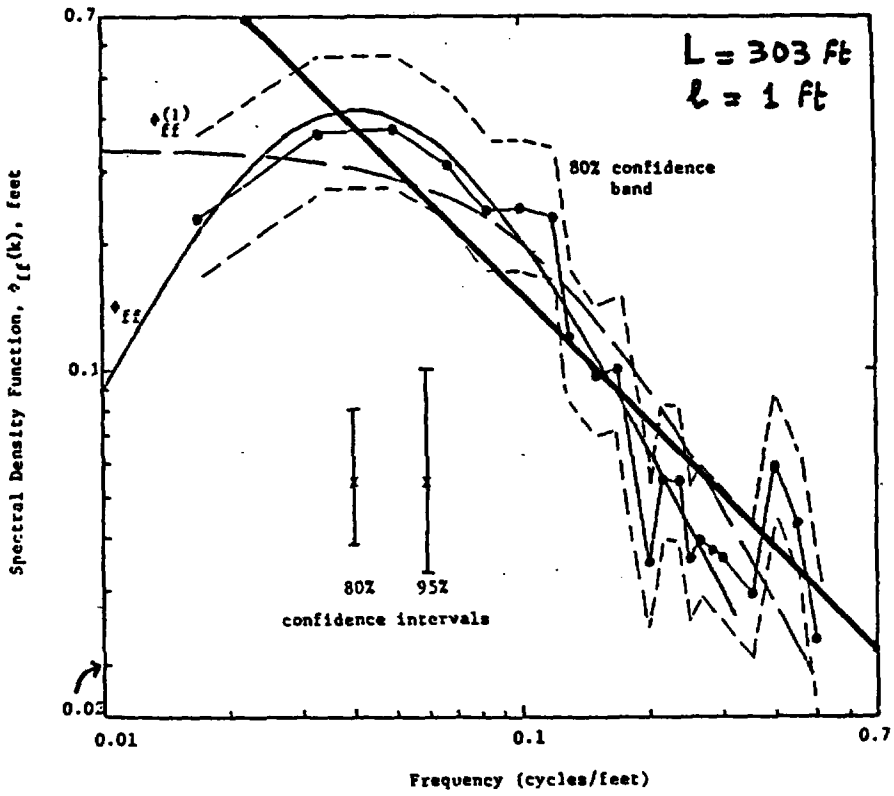


Figure 4.4a: Measured one-dimensional spectrum of log-conductivity at a borehole (Circles), in the Mt. Simon aquifer, from Bakr (1976). The straight line corresponds to a self-similar spectrum with exponent $\alpha = 1$.

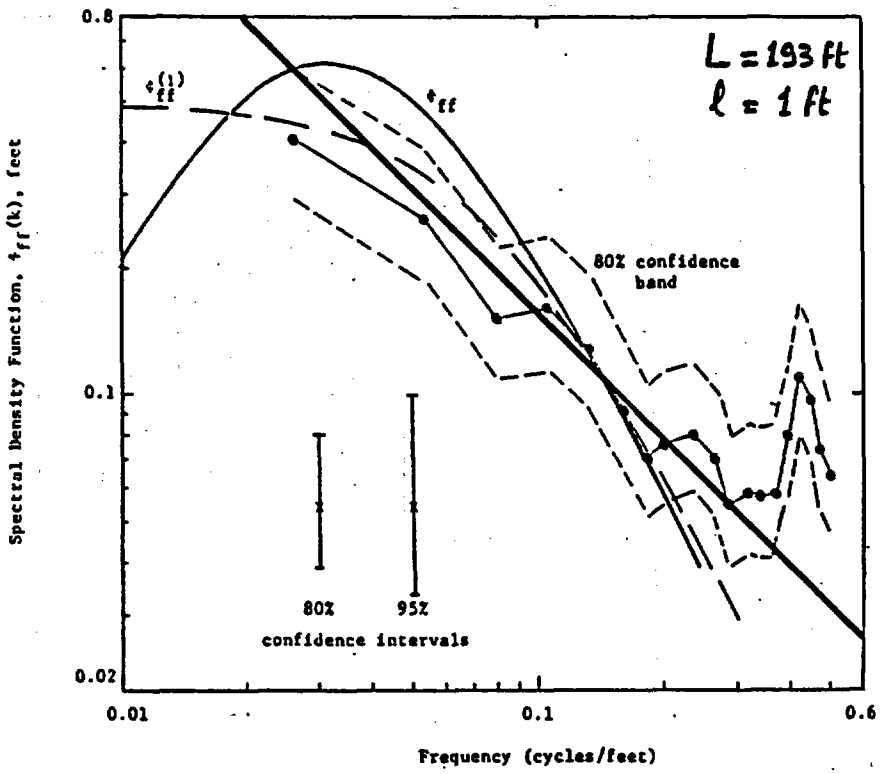


Figure 4.4.b: Same as Figure 4.4.a, for another set of data. The straight line corresponds to a self-similar spectrum with exponent $\alpha = 1$.

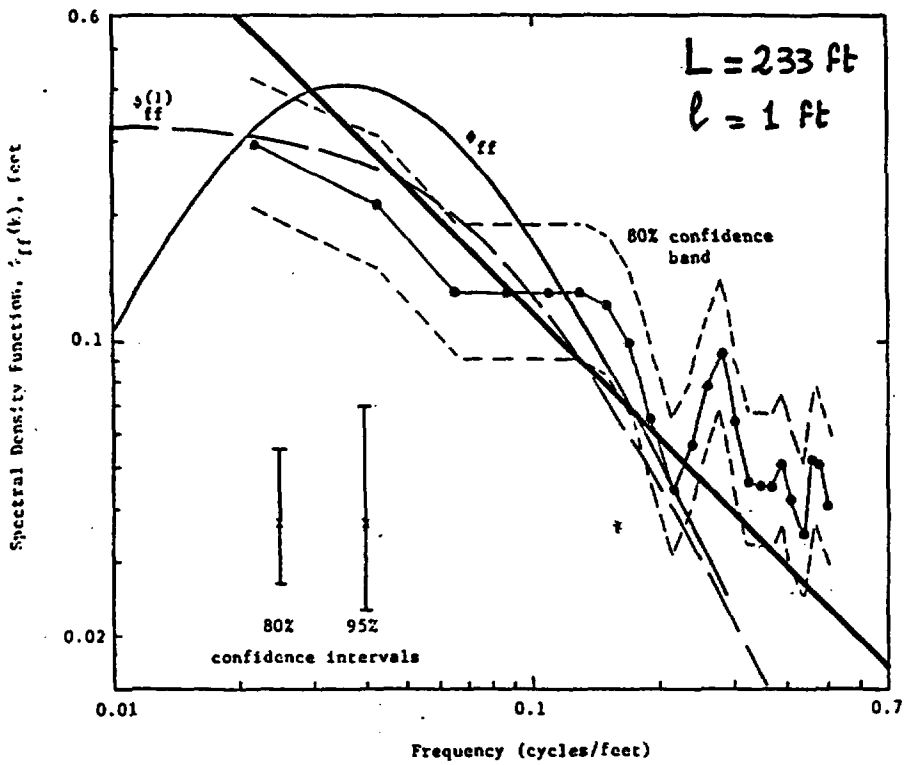


Figure 4.4.c: Same as Figure 4.4a and 4.4b, for another set of data. The straight line corresponds to a self-similar spectrum with exponent $\alpha = 1$.

$$S_{ff}(k) = \frac{S_o}{k^{\alpha}} , \alpha \simeq 1 . \qquad (4.66)$$

It may be instructive at this point to briefly analyze the concept of self-similarity, and the meaning of the coefficient α (slope of the spectrum in a log-log plot). Self-similarity, or fractal behavior, can be thought of as the property of certain physical phenomena to replicate themselves on many scales. This means that certain identifiable fluctuations, structures, patterns, etc., appear to replicate themselves at different length scales, provided a simple similarity transformation. Mandelbrot (1983) exposed the theory of deterministic and *random* fractal geometry, and investigated its manifestations in nature. In particular, Mandelbrot proposed to use the Fractional Brownian Motion (FBM) as a random fractal model of landscapes, and in other applications including hydrological time series and geophysical spatial series (Mandelbrot and Wallis, 1969). More recent studies along these lines include Mandelbrot et al. 1984 (metal fracture surfaces), and Burrough 1981 (spatial analysis of soil granulometry).

The Fractional Brownian Motion is a special class of stationary-increment random process possessing the property of self-similarity (more appropriately "self-affinity"). That is, an FBM process f(x) is self-similar if its increments are statistically invariant under the transformation

$\Delta f(x) \rightarrow \lambda^{-H}\Delta f(\lambda x)$. The parameter H is the Hurst coefficient (or Hölder coefficient, as Δf satisfies a Hölder condition).

A complete review of the properties of one-dimensional FBM processes can be found in Mandelbrot and Van Ness (1968). Note in particular that the usual Brownian motion corresponds to $H = 1/2$. The FBM processes obtained for $1/2 < H < 1$ are "persistent" (positively correlated increments), while $0 < H < 1/2$ gives "antipersistent" processes with negatively correlated increments. The spectrum of the FBM is precisely that of equation (4.66) with $\alpha = 2H + 1$. Observe that the Hurst coefficient is close to zero for the log-conductivity spectra shown in Figures (4.4), indicating a very noisy, anticorrelated behavior within the range of scales available to us (1 ft - 300 ft). This finding seems in accordance with the survey by Burrough (1981), who obtained a low Hurst coefficient for sand and clay fractions ($H \simeq 0.2$).

Incidentally, note that the Hurst coefficient could be evaluated directly from sample functions in physical space rather than spectral densities in Fourier space. Indeed, for an FBM process, the "core function" (variance of increments) takes the form:

$$C_H(\xi) \sim |\xi|^{2H}$$

where ξ is the lag distance. For the data of Figure 4.4, the core function would be nearly flat since $H \simeq 0$. However, this would need to be checked directly by plotting the core function without detrending the log-conductivity data. Other methods, such as "R/S analysis", could also be applied: see Mandelbrot (1983) for reference.

At any rate, the available data shown in Figures (4.4) do indicate that the self-similar spectrum (4.66) with exponent one is a reasonable model, at least within the range of length scales 1 - 300 ft. One should keep in mind that the spectral densities obtained at wavenumbers on the order of inverse domain size or below are unreliable. In our opinion, even the estimated confidence intervals (dashed lines in Figures 4.4) cannot be trusted in the low wavenumber region of the spectrum.

According to the previous discussion, the one-dimensional spectrum of log-conductivity can be approximated as a band-pass self-similar spectrum with Hurst coefficient zero:

$$S_{ff}^{(1)}(k_1) = \begin{cases} S_0/|k_1| & \text{for } L^{-1} \lesssim |k_1| \lesssim \ell^{-1} \\ 0 & \text{otherwise} \end{cases} . \qquad (4.67)$$

Furthermore, we now adopt the view that the large scale cut-off L represents the size of the region in which flow or transport takes place (which may evolve in time), while ℓ represents some

fixed small scale on the order of measurement spacing or perhaps sample size. The intercept of the spectrum at the origin (in a log-log plot) could be thought of as an intrinsic measure of variability of the formation independent of the particular cut-off scales L and ℓ. Accordingly, we define the "intrinsic variance" of log-conductivity as:

$$\sigma_0^2 = S_0. \tag{4.68}$$

On the other hand, the "observed" variance of the log-conductivity within the range of scales (ℓ,L) can be computed by integrating the 1D band-pass spectrum as follows:

$$\sigma_f^2 = 2 \int_{L^{-1}}^{\ell^{-1}} \frac{\sigma_0^2}{k_1} dk_1 = 2\sigma_0^2 \ell n(\frac{L}{\ell}). \tag{4.69}$$

Note that σ_f^2 is scale-dependent, being a slowly varying function of the ratio (L/ℓ). In particular, it is interesting to note that the "observed" variance of the log-conductivity increases logarithmically with domain size.

We now generalize this approach for the three-dimensional case, assuming that the log-conductivity is *statistically isotropic*. First, note that the one-dimensional spectra displayed in Figure 4.4 were obtained by Fourier

transforming the log-conductivity data sampled along a single direction. This yields a one-dimensional marginal spectrum, related to the full 3D spectrum by:

$$S_{ff}^{(1)}(k_1) = \iint S_{ff}(k_1,k_2,k_3)\, dk_2\, dk_3$$

Let us now define a three-dimensional isotropic band-pass self-similar spectrum analogous to its one-dimensional counterpart:

$$S_{ff}(\underline{k}) = \begin{cases} \dfrac{\sigma_0^2}{2\pi k^3} & \text{for } L^{-1} \lesssim k \lesssim \ell^{-1} \\ 0 & \text{otherwise} \end{cases} \qquad (4.70)$$

where k is the radial wavenumber:

$$k = \sqrt{k_1^2+k_2^2+k_3^2}.$$

By plugging equation (4.70) into the equation preceeding it, one obtains the one-dimensional marginal spectrum corresponding to (4.70). It turns out that this spectrum approximates quite well the one-dimensional spectrum of (4.67), at least far enough from the cut-offs, as shown below:

$$S_1(k_1) = \sigma_0^2 \left[\frac{1}{(k_1^2+L^{-2})^{1/2}} - \frac{1}{(k_1^2+\ell^{-2})^{1/2}} \right] \simeq \frac{\sigma_0^2}{|k_1|} \text{ for } L^{-1} \ll k_1 \ll \ell^{-1}.$$

Since the above spectrum agrees fairly well with the 1D spectral data displayed in Figure 4.4, this justifies our choice of the 3D spectrum (4.70) -- assuming only that the formation is statistically isotropic. The more general case of anisotropic media will not be discussed here.

To complete this preliminary investigation, note that the "observed" finite-domain variance of the log-conductivity in 3D space can be obtained by integrating the isotropic spectrum (4.70), leading to the same relation as in the 1D case:

$$\sigma_f^2 = 2\sigma_o^2 \, \ell n \, (L/\ell) \qquad (4.71)$$

where L and ℓ are now the *radial cut-off scales* in 3D space. Again, note that σ_f^2 increases (logarithmically) with domain size. We are now ready to investigate the stochastic flow problem with the one-dimensional and three-dimensional band-pass self-similar spectra defined above.

4.4.3 Stochastic flow solutions for band-pass self-similar spectra:

We now proceed to develop first order spectral solutions following the method already used in Chapter 3. It is important to note that the log-conductivity fields having band-pass

self-similar spectra (4.67) or (4.70) are, by construction, statistically homogeneous in infinite space. As in Chapter 3, we assume again that the mean log-conductivity is a uniquely defined constant, as well as the mean hydraulic gradient. However, we must emphasize the fact that these assumptions do not take into account the influence of domain size on the "observed" mean values. Admittedly, the present approach will be only of limited interest if one insists on stationarity in the mean as a requisite for obtaining tractable solutions. Postponing to a later stage this delicate point, we examine here strictly the solutions obtained by applying the standard first order spectral method with the band-pass self-similar log-conductivity spectra defined above.

For the head variance, the effective conductivity, and the longitudinal macro-dispersion of a convected solute (as defined by Gelhar and Axness, 1983), the calculations are straightforward and need not be detailed here. Recall only that J represents the constant mean hydraulic gradient (expectation or infinite domain average). For one-dimensional flow, with the log-conductivity spectrum (4.67), the head variance is found to be:

$$\boxed{\sigma_h^2 = \sigma_0^2 \, J^2 \, (L^2 - \ell^2)} \qquad (4.72)$$

For three-dimensional flow, with the isotropic spectrum (4.70), the head variance is:

$$\boxed{\sigma_h^2 = \frac{1}{3}\,\sigma_0^2\; J^2(L^2-\ell^2)} \qquad (4.73)$$

In comparison, the results obtained respectively for a 1D hole-exponential model (Gutjahr and Gelhar 1981) and the 3D Markov spectrum were, respectively:

$$\sigma_h^2 = \sigma_f^2\; J^2\; \lambda_{3D}^2 \qquad (4.74)$$

$$\sigma_h^2 = \frac{1}{3}\,\sigma_f^2 J^2 \lambda_{1D}^2. \qquad (4.75)$$

Note that the λ's do not necessarily represent the same length scale in (4.74) and in (4.75). Nevertheless, it is interesting to note that, when $\ell \ll L$ in the band-pass models, then the results obtained with the band-pass and continuous spectra are similar in form if one replaces σ_f^2 by σ_0^2 and λ by L. Since L represents the domain size in (4.72) and (4.73), the head variance obtained with the band-pass spectra appears to depend mainly on large-scale structures, and not on small-scale fluctuations. For instance, assuming $\ell \ll L$ in (4.73) yields the simple result:

$$\sigma_h^2 \simeq \frac{1}{3}\sigma_0^2 J^2 L^2 \qquad (4.76)$$

where σ_0^2 is the "intrinsic" log-conductivity variance and does not depend on the domain size.

For the 3D effective conductivity, one obtains formally the same result as with the 3D Markov spectrum used in Chapter 3, namely:

$$\hat{K}_{11} = K_G \exp(\sigma_f^2/6). \qquad (4.77)$$

However, remember that σ_f^2 now represents the observed finite-domain variance of the log-conductivity within the range of scales (ℓ, L). Accordingly, plugging (4.71) in (4.77) gives a *scale-dependent effective conductivity*:

$$\boxed{\hat{K}_{11} = K_G \cdot \left(\frac{L}{\ell}\right)^{\sigma_0^2/3}} \qquad (4.78)$$

For the data of Figure 4.4.a, the intrinsic log-conductivity variance σ_0^2 is on the order of 0.06, so that the rate of growth of the effective conductivity with domain size is quite slow (+10% for an increase of L by 2 orders of magnitude). Note however that this example corresponds to a mildly variable formation ($\sigma_f^2 = 0.67$ from equation 4.71 with

ℓ = 1 ft and L = 300 ft). For highly variable media, we expect values of σ_0^2 on the order of 0.3 - 0.6. In these cases the effective conductivity (4.78) could increase by a factor of two or more as the domain size L increases by two orders of magnitude.

It is interesting to note that this "size" effect is captured by (4.78) based solely on the standard spectral theory of *stationary* flow fields previously developed in Chapter 3. The increase of effective conductivity with domain size can be interpreted as follows. When the size of the domain of interest is allowed to grow, larger heterogeneities are included. The high conductivity zones outweight those of low conductivity, due to the high degree of freedom of fluid trajectories in a three-dimensional isotropic medium.

In one dimension, the situation is reversed. Indeed, the one-dimensional effective conductivity is the harmonic mean:

$$\hat{K}_{11} = K_H = K_G \exp(-\sigma_f^2/2). \qquad (4.79)$$

Expressing, as before, the observed variance σ_f^2 in terms of the intrinsic variance σ_0^2, we obtain the 1D effective conductivity:

$$\boxed{\hat{K}_{11} = K_G \cdot \left(\frac{L}{\ell}\right)^{-\sigma_0^2}}. \qquad (4.80)$$

Thus, for σ_0^2 on the order of 0.3 - 0.6, the effective conductivity of a one-dimensional flow system will *decrease* by one order of magnitude when the domain size is increased by two orders of magnitude. This is due to the low degree of freedom of fluid trajectories in one dimension. As domain size is increased, larger heterogeneities are included, and the low conductivity zones outweight the high conductivity zones (consider what happens if an impervious inclusion is placed in a 1D flow system). Finally, the two-dimensional case appears to be special. The effective conductivity is equal to the geometric mean, and does not depend on domain size, unlike the 1 and 3-dimensional cases.

The same approach can be used to evaluate the macro-dispersivity of a convected plume. Following Gelhar (1987), the longitudinal macro-dispersity *at large times* is given by:

$$A_{11} = \frac{\pi}{\bar{Q}_1^2} \int_{-\infty}^{+\infty}\int Sq_1q_1(0,k_2,k_3)\, dk_2\, dk_3. \qquad (4.81)$$

Using the flux-based flow equation (Section 4.3) and plugging the 3D Band-Pass Self-Similar spectrum eventually leads to:

$$S_{q_1 q_1}(\underline{k}) = \overline{Q_1^2} \frac{\sigma_0^2}{2\pi} \left[1 - \frac{k_1^2}{k^2}\right]^2 \cdot \frac{1}{k^3} \qquad (4.82)$$

with the provision that $S_{q_1 q_1}$ vanishes outside the wavenumber range (L^{-1}, ℓ^{-1}). Plugging (4.82) into (4.81) and integrating gives finally:

$$\boxed{A_{11} = \sigma_0^2 \cdot (L - \ell)}. \qquad (4.83)$$

This last result suggests that the macro-dispersivity could grow linearly with the size of the plume, due to the fact that a wider range of conductivity heterogeneities will contribute to mechanical dispersion as the plume grows. However, equation (4.83) is admittedly oversimplified, due to the "large time" assumption that was used. Nevertheless, the result does suggest that the macro-dispersivity could increase in time as:

$$\boxed{A_{11}(t) \simeq \sigma_0^2 \cdot L(t)} \qquad (4.84)$$

where $L(t)$ is some typical global scale characteristic of the plume. This relation indicates the occurrence of a positive

feedback: a large macro-dispersivity enhances spreading, which in turn yields a larger macro-dispersity.

In summary, the new spectral solutions obtained with band-pass self-similar spectra reveal the influence of domain size. For a given heterogeneous formation, the analysis suggests that the 3D head standard deviation and the longitudinal macrodispersivity increase like domain size, and the 3D effective conductivity increases slowly as a small power of domain size. In one dimension, though, the effective conductivity decreases with domain size. This behavior was found to be physically realistic based on intuitive arguments.

4.4.4 Uncertainty and "spectral conditioning"

We now proceed to develop further the finite-domain approach, building on previous results obtained with the Band-Pass Self-Similar spectrum. Our purpose here is to show how the effects of local trends could be taken into account by using the idea of "spectral conditioning". This will be illustrated summarily for the simple case of one-dimensional flow in a random self-similar medium, with a prescribed uniform mean flow at the regional scale.

The idea of "spectral conditioning" -- to be described shortly -- was borrowed from a similar technique used in the Fourier-Space version of the Renormalization Group Theory of critical phenomena in statistical mechanics and quantum physics (Wilson and Kogut, 1974; Wilson, 1975; Wilson, 1983). Wilson's version of the Renormalization Group Theory involves defining an average state (magnetization) over regions of a given size (L) by truncating a spectral representation to include only the low wavenumber range ($0 \lesssim k \lesssim L^{-1}$), and obtaining *effective properties* (a Hamiltonian) by conditionally averaging over local fluctuations ($k > L^{-1}$). In fact, this is just the first step of the whole renormalization procedure, which involves iterating and re-scaling to advance towards a fixed point solution. The approach proposed here does not involve the whole recursion, but uses the idea of conditioning low frequencies while solving for higher frequencies.

Consider now the one-dimensional stochastic flow equation:

$$\frac{d^2H}{dx^2} + \frac{dF}{dx}\frac{dH}{dx} = 0 \qquad (4.85)$$

where the log-conductivity F(x) is a random field whose spectrum is Band-Pass Self-Similar, with cut-off wavenumbers ($\mathcal{L}^{-1} \lesssim k \lesssim \ell^{-1}$). Suppose we are interested in finding some

statistical characteristics of the flow within a domain of size L comprised between ℓ and $\mathcal{L}$. This is illustrated in Figure (4.5). The small scale ℓ corresponds to the measurement scale or spacing; the intermediate scale L represents the size of the flow domain of interest (perhaps the size of a contaminated zone); and the large scale $\mathcal{L}$ is a regional scale such that the log-conductivity spectrum is roughly self-similar up to fluctuation scales on the order of $\mathcal{L}$. If a piece of size L of the sample function F(x) is isolated as shown in Figure (4.5), it will appear that the mean value $\overline{F}(x)$ estimated over the domain of size L is different from the regional mean $\langle F \rangle$. Our main purpose is to incorporate this kind of effect in the stochastic equation (4.85) by way of "spectral conditioning".

We now develop the spectral conditioning approach to solve equation (4.85). The key to this approach consists in writing the spectral representation of $F(\underline{x})$, a stationary-ergodic random field, in such a way that the wavenumbers above and below L^{-1} are clearly distinguished:

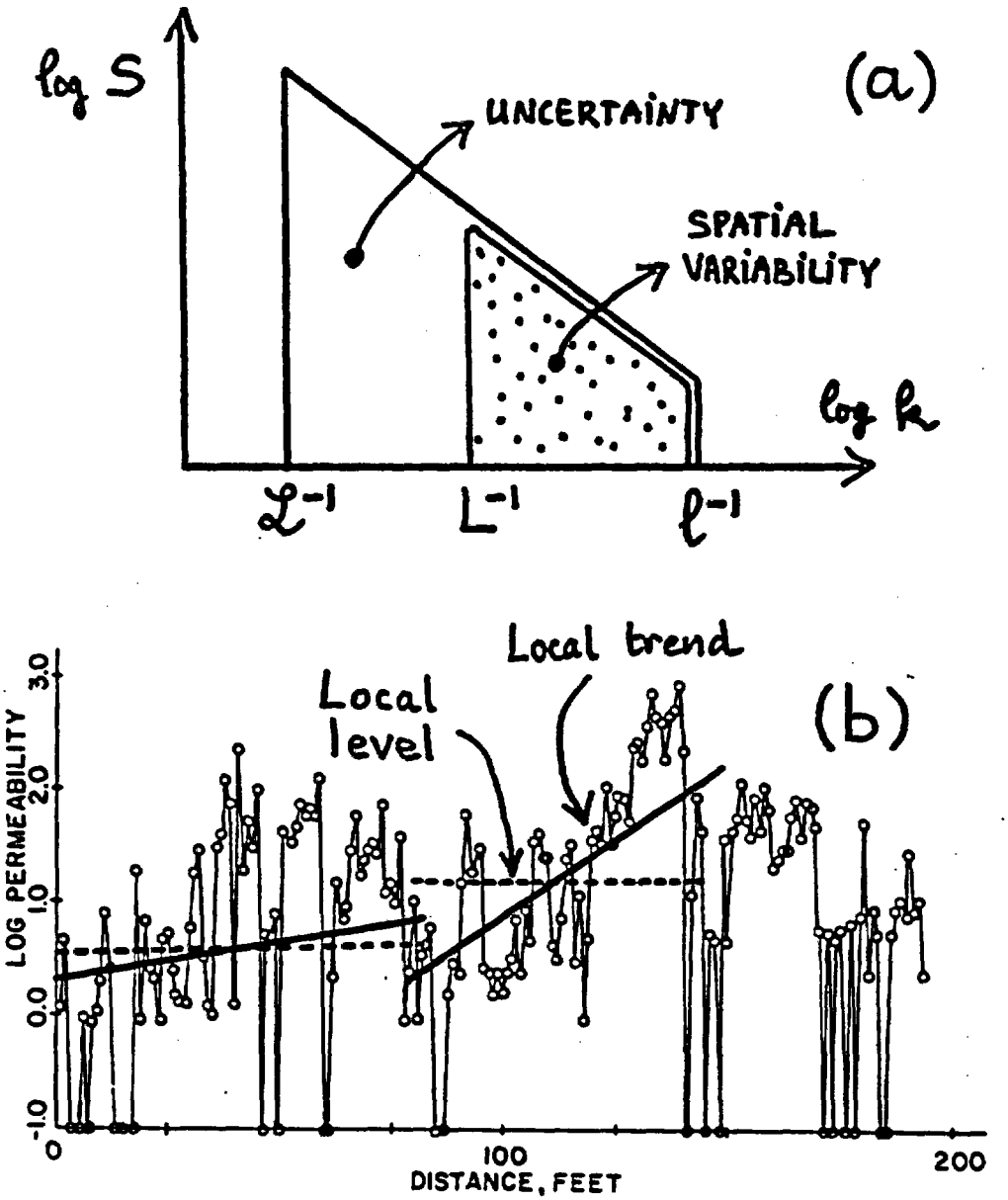


Figure 4.5: Illustration of the spectral conditioning method:
(a) Spectral density versus wavenumber on a log-log plot
(b) Sample function of log-conductivity in space (Mt. Simon data).

$$F(x) = \langle F \rangle + \bar{f}(x) + \tilde{f}(x) = \langle F \rangle + f(x)$$

$$\bar{f}(x) = \int_{\mathcal{L}^{-1} \leq |k| \leq L^{-1}} e^{jkx}\, dZ_F(k) \qquad (4.86)$$

$$\tilde{f}(x) = \int_{L^{-1} \leq |k| \leq \ell^{-1}} e^{jkx}\, dZ_F(k)$$

In these equations, $\langle F \rangle$ represents the ensemble average of log-conductivity, equivalent to infinite domain average by stationarity-ergodicty of $F(x)$ (in practice, this corresponds to the regional mean). The perturbation $f(x)$ was split in two components. The first component is a smooth trend $\bar{f}(x)$ involving only those fluctuations of scale larger than domain size L, and the second is a rapidly fluctuating component $\tilde{f}(x)$ involving higher wavenumbers. As the size of the domain becomes large (on the order of the regional scale $\mathcal{L}$ or larger) this representation yields the usual infinite-domain spectral representation with:

$$\bar{f}(x) \to 0$$
$$\tilde{f}(x) \to f(x)$$
$$F = \langle F \rangle + f(x).$$

Let us now "freeze" the local trend $\bar{f}(x)$ by assuming that the Fourier coefficients at low wavenumbers are known. This procedure boils down to conditioning the local log-conductivity

field with respect to its large scale fluctuations and to the specific location of the local domain of interest. Say for instance that x_o defines the location of the center of the flow domain:

$$x_o - \frac{L}{2} \leq x \leq x_o + \frac{L}{2} .$$

Thus, $\bar{f}(x)$ is known when the position x_o of the domain and all fluctuations on the order of domain size or larger are known. Plugging (4.86) in (4.85) while holding $\bar{f}(x)$ fixed yields a stochastic equation driven by the local fluctuations $\tilde{f}(x)$, as will be seen shortly.

We now introduce a spectral representation for the head process as well. Using the same model as (4.86) gives:

$$H(x) = \langle H(x) \rangle + \bar{h}(x) + \tilde{h}(x). \qquad (4.87)$$

Plugging (4.86) and (4.87) in (4.85) gives finally the stochastic flow equation in the desired form:

$$\frac{d^2\langle H \rangle}{dx^2} + \frac{d^2\bar{h}}{dx^2} + \frac{d^2\tilde{h}}{dx^2} + \left[\frac{d\bar{f}}{dx} + \frac{d\tilde{f}}{dx}\right]\left[\frac{d\langle H \rangle}{dx} + \frac{d\bar{h}}{dx} + \frac{d\tilde{h}}{dx}\right] = 0 \ . \qquad (4.88)$$

Upon conditional averaging (in the sense described earlier), the conditional fluctuations $\tilde{f}$ and $\tilde{h}$ vanish. This yields a

conditional mean equation:

$$\frac{d^2\langle H\rangle}{dx^2} + \frac{d^2\bar{h}}{dx^2} + \frac{d\bar{f}}{dx}\frac{d\langle H\rangle}{dx} + \frac{d\bar{f}}{dx}\frac{d\bar{h}}{dx} = -\left[\overline{\frac{d\tilde{f}}{dx}\frac{d\tilde{h}}{dx}}\right]. \qquad (4.89)$$

Note again that the "bar" sign stands for a conditional spectral average as defined in (4.86). By subtracting (4.89) from (4.88), one also obtains the equation governing the local head perturbations $\tilde{h}$ conditioned on regional scale fluctuations:

$$\frac{d^2\tilde{h}}{dx^2} + \frac{d\tilde{f}}{dx}\frac{d\langle H\rangle}{dx} + \frac{d\tilde{f}}{dx}\frac{d\bar{h}}{dx} + \frac{d\bar{f}}{dx}\frac{d\tilde{h}}{dx} = -\left\{\frac{d\tilde{f}}{dx}\frac{d\tilde{h}}{dx} - \left[\overline{\frac{d\tilde{f}}{dx}\frac{d\tilde{h}}{dx}}\right]\right\}.$$

Neglecting the second order perturbation term on the right-hand side gives finally:

$$\frac{d^2\tilde{h}}{dx^2} + \frac{d\tilde{f}}{dx}\frac{d\langle H\rangle}{dx} + \frac{d\tilde{f}}{dx}\frac{d\bar{h}}{dx} + \frac{d\bar{f}}{dx}\frac{d\tilde{h}}{dx} \simeq 0. \qquad (4.90)$$

Before attempting to solve (4.90) for $\tilde{h}(x)$, we need some information on $\bar{h}(x)$. This is obtained by solving (4.89) for $\bar{h}(x)$, with $\bar{f}(x)$ considered now as a *random* function (not a prescribed deterministic function). This may be called the "unconditional solution step", where $\bar{f}(x)$ is allowed to vary randomly over all possible fluctuations larger than L, and over all possible locations x_o (midpoint of the local flow domain).

Accordingly, we now apply ensemble averages to equation (4.89) in order to obtain the *unconditional* ensemble mean and the *unconditional* perturbation equations:

$$\frac{d^2\langle H\rangle}{dx^2} = -\left\langle \frac{d\bar{f}}{dx}\frac{d\bar{h}}{dx}\right\rangle - \left\langle \overline{\frac{d\tilde{f}}{dx}\frac{d\tilde{h}}{dx}}\right\rangle$$

$$\frac{d^2\bar{h}}{dx^2} + \frac{d\bar{f}}{dx}\frac{d\langle H\rangle}{dx} = -\left\{\frac{d\bar{f}}{dx}\frac{d\bar{h}}{dx} - \left\langle \frac{d\bar{f}}{dx}\frac{d\bar{h}}{dx}\right\rangle\right\}$$

$$-\left\{\overline{\frac{d\tilde{f}}{dx}\frac{d\tilde{h}}{dx}} - \left\langle \frac{d\tilde{f}}{dx}\frac{d\tilde{h}}{dx}\right\rangle\right\}.$$

We now neglect the high order perturbation terms appearing on the right-hand sides, as we did in the formal "small parameter" expansion developed in section 3.1 of Chapter 3. Thus, we obtain finally:

$$\frac{d^2\langle H\rangle}{dx^2} \simeq 0 \qquad (4.91)$$

and

$$\frac{d^2\bar{h}}{dx^2} + \frac{d\bar{f}}{dx}\frac{d\langle H\rangle}{dx} \simeq 0. \qquad (4.92)$$

Equation (4.91) implies that the ensemble mean (or regional) hydraulic gradient is *constant*:

$$J = - \frac{d\langle H \rangle}{dx} \quad . \tag{4.93}$$

Plugging this into equation (4.92) governing the perturbation of the local mean head, yields the simple result:

$$\frac{d\bar{h}}{dx} \simeq J \cdot \bar{f}(x). \tag{4.94}$$

This relation makes sense intuitively. The term $(- \frac{d\bar{h}}{dx})$ represents the perturbation of the local mean hydraulic gradient with respect to the regional gradient (J); it is positive if the local mean log-conductivity is smaller than the regional mean ($\bar{f}$ negative). Thus, the local mean hydraulic gradient will be higher than the regional mean if the local mean log-conductivity is lower than its regional mean (the term "local" refers to domain size L). This behavior is suggested more directly by the form of the Darcy equation in one-dimensional space: the hydraulic gradient is inversely proportional to the conductivity.

We may now attempt a spectral solution of the *conditional* perturbation equation (4.90), by plugging (4.93) and (4.94) while holding $\bar{f}(x)$ fixed (non-random). This gives the required stochastic equation for the *local* head perturbations:

$$\boxed{\frac{d^2\tilde{h}}{dx^2} + \frac{d\bar{f}}{dx}\frac{d\tilde{h}}{dx} \simeq J \cdot (1-\bar{f}) \cdot \frac{d\tilde{f}}{dx} \quad .} \tag{4.95}$$

It is important to note that $\bar{f}(x)$ is by construction a slowly varying function, compared to the rapid oscillations of $\tilde{f}(x)$. This suggests approximating equation (4.95) by letting $d\bar{f}/dx$ be some average slope, and $\bar{f}$ some average level of $\bar{f}(x)$ within the flow domain, as illustrated in Figure (4.5).

With this provision, equation (4.95) becomes a linear stochastic equation which can be solved by the standard spectral method in Fourier space. The result is given below in terms of Fourier-Stieltjes increments conditioned on the values taken by $\bar{f}$ and $d\bar{f}/dx$:

$$dZ_h^{\sim}(k) \simeq - J\,(1-\bar{f})\,\frac{jk}{k^2 - j\frac{d\bar{f}}{dx}k}\,dZ_f^{\sim}(k). \qquad (4.96)$$

This gives the conditional head spectrum:

$$\boxed{S_{hh}^{\sim\sim}(k) \simeq J^2(1-\bar{f})^2\,\frac{1}{k^2+(\frac{d\bar{f}}{dx})^2}\,S_{ff}^{\sim\sim}(k)} \qquad (4.97)$$

where $S_{ff}^{\sim\sim}(k)$ is the conditional band-pass self-similar spectrum σ_0^2/k, vanishing outside the range $L^{-1} \leq k \leq \ell^{-1}$. Note that the low wavenumber range ($k \leq L^{-1}$) has been incorporated entirely into the terms $\bar{f}$ and $d\bar{f}/dx$, the local mean level and local trend of log-conductivity.

The *conditional* head variance obtains directly from (4.97) by integrating over the wavenumber range (L^{-1},ℓ^{-1}). This gives after some manipulation and replacing $(1-\bar{f})$ by $e^{-\bar{f}}$:

$$\tilde{\sigma}_h^2 \simeq \sigma_0^2 J^2 \frac{e^{-2\bar{f}}}{\left(\frac{d\bar{f}}{dx}\right)^2} \cdot \ell n \left\{ \frac{1+\left(\frac{d\bar{f}}{dx}\right)^2 L^2}{1+\left(\frac{d\bar{f}}{dx}\right)^2 \ell^2} \right\}. \qquad (4.98)$$

Incidentally, it is interesting to note that this expression gives the correct head variance as L goes to infinity ($L>>\ell$). Indeed in this case, the local and regional means coincide, yielding $\bar{f} \to 0$ and $\frac{d\bar{f}}{dx} \to 0$, and it can be seen from a Taylor development of the logarithmic term that the right-hand side of (4.98) will coincide exactly with the previous result (4.72). Recall also that the *conditional* log-conductivity variance, obtained by integrating the log-conductivity spectrum in the range $L^{-1} \leq k \leq \ell^{-1}$, depends on domain size as in equation (4.69), that is:

$$\tilde{\sigma}_f^2 = 2\,\sigma_0^2\,\ell n\left(\frac{L}{\ell}\right). \qquad (4.99)$$

The last step of the spectral conditioning method consists in analyzing the local head variance (4.98) as a random parameter when $\bar{f}(x)$ is viewed as a random field. Indeed, remember that $\bar{f}$ was defined in (4.86) as the difference between

a local and regional mean: the spectral content of $\overline{f}(x)$ coincides with the spectrum of the log-conductivity in the wavenumber range $\mathcal{L}^{-1} \leq k \leq L^{-1}$. Accordingly, $\tilde{\sigma}_h^2$ can be viewed as the local head variance due to spatial variability at the local scale L; at the same time, $\tilde{\sigma}_h^2$ also appears to be random at the regional scale $\mathcal{L}$. The randomness of $\tilde{\sigma}_h^2$ arises from uncertainty among all possible realizations of the regional formation, and should be understood in the Bayesian sense, that is in terms of *risk analysis*.

In order to evaluate explicitly the uncertainty on $\tilde{\sigma}_h^2$, let us compute the ensemble variance of the random terms appearing in (4.98) for the case $L \leq \mathcal{L}$:

$$\mathrm{Var}(\overline{f}) = 2\sigma_0^2 \, \ell n(\frac{\mathcal{L}}{L})$$

$$\mathrm{Var}(\frac{d\overline{f}}{dx}) = \sigma_0^2 \quad (\frac{1}{L^2} - \frac{1}{\mathcal{L}^2}) \; .$$

Plugging these values into (4.98), we obtained after some manipulations a rough estimate of the coefficient of variation of the local head variance:

$$C.V.(\tilde{\sigma}_h^2) \sim 1 - \exp\{-2\sqrt{2\sigma_0^2 \ell n \; (1/\epsilon)}\} \cdot \frac{\ell n\{1+\sigma_0^2(1-\epsilon^2)\}}{\sigma_0^2(1-\epsilon^2)} \qquad (4.100)$$

where ϵ is the length scale ratio L/ℓ. Using hypothetical data for a fairly heterogeneous formation ($\sigma_0^2 \simeq 0.5$) with a domain size about one order of magnitude smaller than the regional scale ($\epsilon = 0.1$), yields a C.V. on the order of 100%. In addition, equation (4.98) shows more specifically that the local head variance is larger than its expectation if the local mean of the log-conductivity is smaller than the regional mean value.

A similar analysis can be carried out for the *effective conductivity*. First, the Darcy equation can be used to define a *local* effective conductivity $\tilde{K}_{eff}$, conditioned on regional scale fluctuations. Second, the *uncertainty* on $\tilde{K}_{eff}$ can be computed by averaging over the unresolved fluctuations (terms involving $\overline{f}(x)$). Using previous results (equation 4.94), the Darcy equation can be expressed as:

$$Q = - K_G \, e^{(\overline{f}+\tilde{f})} \cdot (- J + J \cdot \overline{f} + \frac{\tilde{dh}}{dx})$$

where the flux Q is necessarily constant for 1D flow. Taking conditional expectations with respect to both sides (with $\overline{f}(x)$ fixed), and linearizing the exponential term gives:

$$Q \simeq - K_G \, \{J \cdot (\frac{\overline{f}^2}{2} - 1) + \overline{\tilde{f} \cdot \frac{\tilde{dh}}{dx}} - J \cdot \frac{\overline{(\tilde{f}^2)}}{2}\}.$$

Using previous definitions, the second term on the right can be computed by spectral integration as shown below:

$$\overline{\tilde{f}\,\frac{dh}{dx}} = \int_{L^{-1}\leq|k|\leq\ell^{-1}} jk\,\langle dZ_f^{*}\cdot dZ_h\rangle .$$

Plugging equation (4.96) for dZ_h, and using the band-pass self-similar spectrum for the log-conductivity finally leads to:

$$Q = \tilde{K}_{eff}\cdot J$$

where the local effective conductivity is given by:

$$\tilde{K}_{eff} \simeq K_G\cdot\left\{1-\frac{\overline{f^2}}{2}+\sigma_0^2\,\ell n\left(\frac{L}{\ell}\right)\right.$$

$$\left.-(1-\bar{f})\cdot 2\sigma_0^2\,\ell n\left[\frac{L}{\ell}\cdot\sqrt{\frac{1+a^2\ell^2}{1+a^2L^2}}\right]\right\}$$

and $a = d\bar{f}/dx$. After rearranging and replacing terms like $(1-\bar{f})$ by exponentials, we obtain a more realistic expression for the local effective conductivity as follows:

$$\tilde{K}_{eff} \simeq$$

$$K_G \cdot \exp\left\{(1-2e^{-\bar{f}})\cdot\sigma_o^2 \ell n(\frac{L}{\ell}) + e^{-\bar{f}^2/2} + e^{-\bar{f}}\sigma_o^2 \ell n\left[\frac{1+a^2L^2}{1+a^2\ell^2}\right]\right\}. \quad (4.101)$$

This effective conductivity is local to the flow domain of size L, in the sense that it is conditioned on the local mean ($K_G e^{\bar{f}}$) and the local trend ($a = d\bar{f}/dx$). As the domain size increases, these parameters converge to the regional mean values ($\bar{f} = 0$ and $a = 0$). In that case, equation (4.101) converges to the harmonic mean:

$$K_H = K_G e^{-\sigma_o^2 \ell n(\frac{\ell}{\ell})} = K_G e^{-\sigma_f^2/2}$$

as expected. Moreover, equation (4.101) shows that the local effective conductivity drops below the arithmetic mean if the local mean level of $\ell n\, K$ happens to be smaller than the regional mean ($\bar{f} < 0$). Finally, the variance of $\tilde{K}_{eff}$ could be computed from this equation by letting $\bar{f}(x)$ be random, and applying ensemble averages to resolve the uncertainty due to low wavenumber components (regional scale fluctuations).

In summary, the "spectral conditioning" approach based on a band-pass self-similar model of the log-conductivity spectrum was used to evaluate the effect of domain size on "large scale" flow characteristics. This led to closed form expressions for the head variance and the effective conductivity in a *finite domain*. Moreover, we have shown that these statistical quantities were themselves subject to *uncertainty* due to heterogeneities on the order of domain size or larger. The solutions obtained for the case of one-dimensional flow incorporate such uncertainty in the form of two simple random parameters:

(i) The local mean "level" of conductivity ($K_G e^{\overline{f}}$);

(ii) The local mean slope of log-conductivity ($a=d\overline{f}/dx$).

These local parameters were defined by smoothing out the local fluctuations of $\ell n\, K$, i.e. those fluctuations whose wavenumber is higher than the inverse domain size. Their spectral content is concentrated exclusively in the low wavenumber range, which decreases as domain size increases.

The whole approach leads to analytical results that distinguish the uncertainty and spatial variability of phenomena taking place over finite scales. This may be particularly relevant for the case of a subsurface contamination plume

spreading from a local source, where there is some uncertainty about the location of the center of mass and the extent (macro-dispersion) of the plume at early times. It would be interesting to examine how this uncertainty decreases as the time and length scales of the plume increase. Admittedly, more work is needed in order to evaluate the potential of the proposed "spectral conditioning approach" for realistic model problems of stochastic flow and dispersive transport. Last but not least, large sets of conductivity data collected over a wide range of scales may be needed in order to ascertain the validity of the proposed band-pass self-similar model. A useful range of scales for subsurface contamination problems may involve at least three to four orders of magnitude along each spatial direction.

CHAPTER 5 NUMERICAL METHOD FOR LARGE SINGLE-REALIZATION SOLUTIONS OF STOCHASTIC FLOW IN SATURATED OR UNSATURATED MEDIA

5.1 Governing Equations and Finite Difference Approximations

In this introductory section, we develop a Finite Difference numerical approximation for solving *large* single-realization stochastic flow problems in three dimensions. The choice of the finite difference (FD) discretization method is justified by considering the numerical requirements and computational work involved. The discrete form of the flow equation (finite difference system) is given in detail for steady saturated flow, and also for the more general case of transient flow in partially saturated or unsaturated media with nonlinear and spatially variable coefficients.

5.1.1 - Governing equation and numerical requirements

For simplicity, let us focus here on the case of saturated flow in the steady state, postponing to a later stage the analysis of the more general case of transient and unsaturated flow. The governing equation for the hydraulic head H is easily obtained from the Darcy equation and the continuity equation. By using implicit summation over repeated indices, this yields in three dimensions:

$$L_K(H) = -\frac{\partial}{\partial x_i}\left[K(x)\frac{\partial H}{\partial x_i}\right] = 0 \qquad (i = 1,2,3). \qquad (5.1)$$

Here, we emphasize once again that the flow problem to be solved is inherently *stochastic*, since the conductivity is assumed to be a random function of space. But, according to the *single realization approach*, we aim at obtaining a numerical solution of the flow equation for one particular realization of the random conductivity field $K(\underline{x})$, over a large finite domain with specified boundary conditions. Thus, equation (5.1) is now considered as a single realization of a stochastic partial differential equation. The flow problem "appears" to be deterministic and may be solved, in principle, by standard numerical methods.

On the other hand, recall that the idea of the single-realization approach is to obtain the *statistical* properties of the flow field by postulating the equivalence of spatial averages and ensemble averages. The flow statistics obtained in this manner should be unique, independent of the particular realization, provided that the flow field is statistically homogeneous and ergodic and the domain is sufficiently large. In this framework, the numerical solution of the single-realization problem requires special care regarding the numerical method to be used. The difficulty is that the local conductivity $K(\underline{x})$ fluctuates wildly in space, which

requires a high resolution mesh in order to capture the detailed features of the flow field. On the other hand, the domain must be large compared to the largest scale of fluctuation of the flow field, in order to guarantee the equivalence of spatial and ensemble statistics by the ergodicity hypothesis. These considerations imply that the size of the computational grid may have to be unusually large to capture the fluctuations of the flow solution over a reasonably wide range of scales.

For a preliminary evaluation of statistical requirements, one may use the ℓn K-correlation scale λ_i as an indication of the typical scale of fluctuation of the solution (the index i refers to the spatial direction x_i). Let Δx_i be the size of the discretization cell, and L_i the size of the computational flow domain. We require for adequate statistical resolution:

$$\Delta x_i / \lambda_i \ll 1. \tag{5.2}$$

On the other hand, the computational flow domain must be large enough so that many "independent events" can be sampled. In other words, the domain size must be taken much larger than the scale of fluctuation, that is:

$$L_i / \lambda_i \gg 1. \tag{5.3}$$

Now, the number of discretization cells in each direction is:

$$n_i = L_i/\Delta x_i$$

and the total number of cells in 3 dimensions is:

$$N = n_1 n_2 n_3.$$

The, requirements (5.2) and (5.3) could lead to ratios $L_i/\Delta x_i$ on the order of 100 or more, implying that the typical size of the computational grid (N) could be 1 Million cells or more.

Another major point of concern is the validity of classical numerical methods, such as finite differences (FD) or finite elements (FE) Methods, when applied to an equation like (5.1) with highly variable conductivities. Intuitively, one would expect that the error due to discretization decreases as σ_f decreases. Note for instance that as $\sigma_f \to 0$, equation (5.1) becomes just the Laplace equation which can be solved accurately (even exactly) with a second order accurate FD or FEM scheme. The key question is: "what happens when σ_f is significantly different from zero?" The truncation error analysis to be developed in the next section will show that the discretization error decreases with $\Delta x_i/\lambda_i$ when σ_f is fixed. Intuitively, this means that truncation errors are small when the solution

behaves smoothly at the scale of the mesh. We conclude that the resolution constraint (5.2) is a requirement for *numerical accuracy* as well as statistical resolution.

In summary, the issues of numerical accuracy, statistical resolution, and large sampling space, all indicate the need for solving very large discrete systems (e.g., one million equations or more). This has led us to the choice of the 7-point centered finite difference scheme for spatial discretization. The reason for this choice will appear more clearly in the sequel. Let us briefly mention some of the arguments in favor of this discretization method. First of all, the resulting system of equations is very sparse, symmetric and can be solved efficiently by fast iterative methods based on matrix preconditioning, such as the Strongly Implicit Procedure (SIP) and the Incomplete-Choleski Conjugate Gradient (ICCG) methods. Moreover, the algebraic properties of the coefficient matrix arising from other discretization methods, such as Galerkin, would not be as well suited to fast solution methods. In our view, the centered finite difference scheme essentially gives the most sparse and best structured algebraic system among all discretization methods consistent with the governing flow equation. Based on these remarks, most of our efforts were devoted to developing efficient solvers for large finite difference systems having spatially variable (random)

coefficients. This will be developed in detail in Section 5.3 (for linear systems) and in Section 5.4 (for unsaturated nonlinear systems), following the truncation error analysis of linear flow systems presented in Section 5.2.

We now proceed to develop the finite difference approximations, first for the steady saturated flow equation, then for the transient unsaturated flow equation with nonlinear coefficients. The stochastic nature of the coefficients in these equations should always be kept in mind, particularly for purposes of error analysis.

5.1.2 - Finite difference in space for steady saturated flow

[a] Derivation of the Finite Difference System:

The 7-point centered finite difference approximation of the steady state flow equation (5.1) obtains by approximating the Darcy flux $q_i = - K(\underline{x})\frac{\partial H}{\partial x_i}$ by the centered difference scheme:

$$q_i(x_{i+\frac{1}{2},j,k}) \simeq -K_{i+\frac{1}{2},j,k}\left[\frac{H_{i+1,j,k} - H_{i,j,k}}{\Delta x_i}\right] \qquad (5.4)$$

Note that q_i is evaluated at the mid-nodal point $x_{i+\frac{1}{2},j,k}$ of an orthogonal grid where nodes are located at $x_{i,j,k}$. The derivative $\frac{\partial q_i}{\partial x_i}$ is then approximated again by the centered

difference:

$$\frac{\partial q_1}{\partial x_1}(x_{i,j,k}) \simeq \frac{q_1(x_{i+\frac{1}{2},j,k}) - q_1(x_{i-\frac{1}{2},j,k})}{\Delta x_1} \qquad (5.5)$$

at the node points $x_{i,j,k}$. By repeating similar approximations for the terms $\frac{\partial q_2}{\partial x_2}$ and $\frac{\partial q_3}{\partial x_2}$, one obtains the Finite Difference system governing the hydraulic head at the nodes of the orthogonal grid. For convenience, we use the triple-index notation as shown below between brackets:

$$[0] = (i,j,k)$$

$$[i \pm \tfrac{1}{2}] = (i \pm \tfrac{1}{2},j,k)$$

$$[i \pm 1] = (i \pm 1,j,k)$$

etc.

Accordingly, the FD system for heads can be expressed as:

$$\hat{L}_K(H) = -\frac{K[i-\frac{1}{2}]}{(\Delta x_i)^2} \cdot H[i-1] - \frac{K[j-\frac{1}{2}]}{(\Delta x_2)^2} \cdot H[j-1] - \frac{K[k-\frac{1}{2}]}{(\Delta x_3)^2} H[k-1]$$

$$+ \left\{ \frac{K[i-\frac{1}{2}]+K[i+\frac{1}{2}]}{(\Delta x_1)^2} + \frac{K[j-\frac{1}{2}]+K[j+\frac{1}{2}]}{(\Delta x_2)^2} + \frac{K[k-\frac{1}{2}]+K[k+\frac{1}{2}]}{(\Delta x_3)^2} \right\} \cdot H[0]$$

$$- \frac{K[i+\frac{1}{2}]}{(\Delta x_1)^2} \cdot H[i+1] - \frac{K[j+\frac{1}{2}]}{(\Delta x_2)^2} \cdot H[j+1] - \frac{K[k+\frac{1}{2}]}{(\Delta x_3)^2} \cdot H[k+1] = 0 \qquad (5.6)$$

and the Darcy flux vector (5.4) can be expressed in terms of the discrete head solution as:

$$q_1[i+\tfrac{1}{2}] = -K[i+\tfrac{1}{2}] \cdot \frac{H[i+1]-H[0]}{\Delta x_1}$$

$$q_2[j+\tfrac{1}{2}] = -K[j+\tfrac{1}{2}] \cdot \frac{H[j+1]-H[0]}{\Delta x_2} \qquad (5.7)$$

$$q_3[k+\tfrac{1}{2}] = -K[k+\tfrac{1}{2}] \cdot \frac{H[k+1]-H[0]}{\Delta x_3}$$

where H[0] stands for the head at the central node H(i,j,k).

The FD system (5.6) comprises N equations: one equation per node. Equation (i,j,k) relates the unknown at node (i,j,k) to the unknown at six neighboring nodes (i±1, j±1, k±1) as illustrated in Figure 5.1, representing the 7-point FD molecule in space. Note that mid-nodal conductivities like K[i+½] stand for point values of $K(\underline{x})$ at $\underline{x} = x_{i+\frac{1}{2},j,k}$. Since only the nodal values of $K(\underline{x})$ are known, the mid-nodal conductivities must be approximated by some weighting scheme, such as the *geometric mean of nodal conductivities*:

$$\hat{K}[i+\tfrac{1}{2}] \simeq \sqrt{K[0] \cdot K[i+1]}$$

or in more explicit triple-index notation:

$$\hat{K}(i+\tfrac{1}{2},j,h) \simeq \sqrt{K(i,j,k) \cdot K(i+1,j,k)} \qquad (5.8)$$

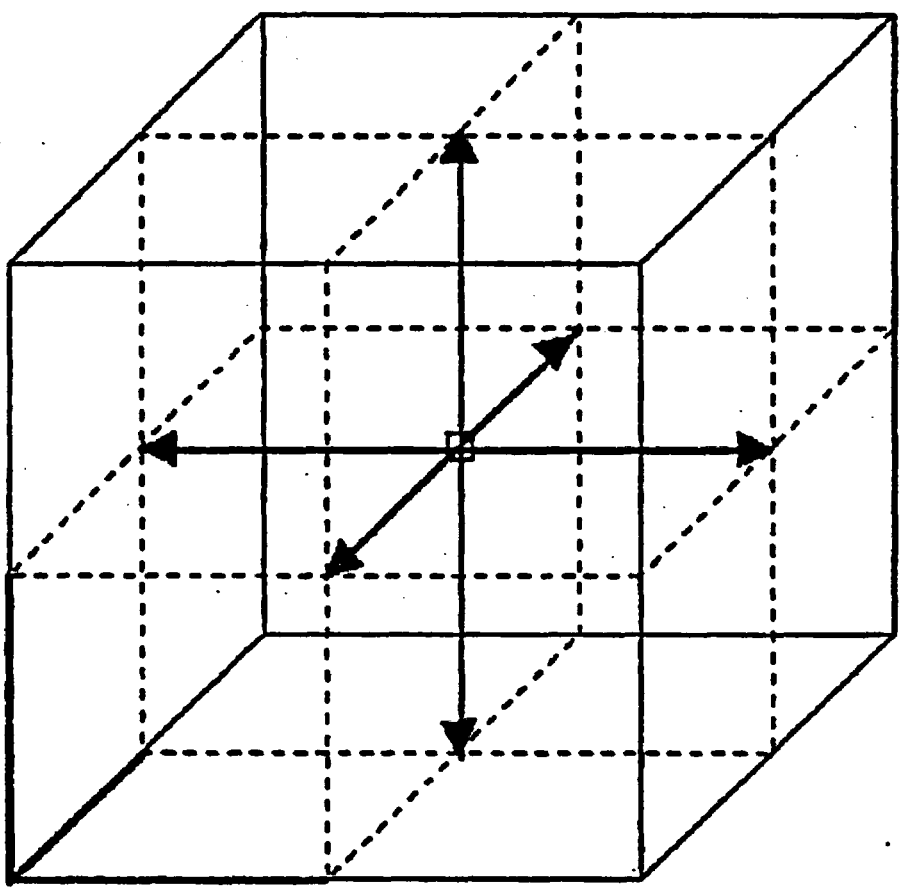


Figure 5.1 Seven-point centered finite difference molecule for a three-dimensional orthogonal grid of mesh points.

The geometric mean weighting scheme was preferred to other schemes used in the literature (such as the arithmetic mean and the harmonic mean) based on the findings of the spectral theory concerning the effective conductivity of a three-dimensional random isotropic formation:

$$K_H = K_G e^{-\sigma_f^2/2} < K_{eff} = K_G e^{\sigma_f^2/6} < K_A = K_G e^{\sigma_f^2/2} \qquad (5.9)$$

Equation (5.9) shows that the flow at the large scale is governed by K_G more closely than K_A or K_H. This suggests that the flow at smaller scales is similarly governed by the local geometric mean conductivity (5.8), more closely than by arithmetic or harmonic averages. To complete this analogy, observe that the discrete conductivity field will appear more nearly isotropic at the mesh scale if the mesh size is chosen proportional to the conductivity correlation scales in all three directions ($\Delta x_1/\lambda_2 = \Delta x_2/\lambda_2 = \Delta x_3/\lambda_3$). This seems to be a desirable property in order to avoid artificial grid-induced anisotropy.

Other criteria for a "best" conductivity weighting scheme were proposed in the literature (e.g., Narasimhan et al., 1978). Their arguments do not appear convincing enough to be taken into account, as truncation error analysis shows that the

weighting error $\hat{K}[i+\frac{1}{2}]-K[i+\frac{1}{2}]$ remains of order $O((\Delta x/\lambda)^2)$ for all consistent weighting schemes such as K_H, K_G, K_A. Intuitively, when the mesh size is such that $\Delta x_i/\lambda_i$ is the same in all directions, the geometric mean weighting scheme appears as a good compromise between the arithmetic mean (exact for layered systems with flow parallel to layers) and the harmonic mean (exact for layered systems with flow perpendicular to layers).

We now focus on the structure of the linear FD system (5.6). In matrix notation, this system can be expressed as:

$$\underline{\underline{K}} \cdot \underline{h} = \underline{b} \qquad (5.10)$$

where $\underline{\underline{K}}$ is the coefficient matrix, or conductivity matrix of size $N = n_1 n_2 n_3$ for the 3D case (boundary nodes excluded). The vector $\underline{h}$ represents the nodal head values, and $\underline{b}$ is a vector containing boundary conditions such as fixed head and fixed flux. In our implementation, this vector was formed by a technique known as "matrix condensation": the discretized boundary conditions were used to express the unknown at the boundaries in terms of the unknowns at neighboring nodes located inside the domain. All quantities such (as H or q) specified at the boundaries were then transferred to the right-hand side of the system (vector $\underline{b}$). The details of this procedure are illustrated below in the simple case of one-dimensional flow.

Consider the case of one-dimensional flow with fixed head on the left boundary and fixed flux on the right boundary. The flux condition on the right is discretized by using a centered FD approximation at the mid-node located between the last and next-to-last nodes (one may consider the physical boundary to be located precisely at the mid-node). The finite difference approximation of Darcy equation yields:

$$q(x_{n+\frac{1}{2}}) \simeq -K_{n+\frac{1}{2}} \cdot \frac{H_{n+1}-H_n}{\Delta x} = q_{n+\frac{1}{2}} \qquad (5.11)$$

On the other hand, the head condition on the left boundary can be expressed exactly as:

$$H(x_o) = H_o. \qquad (5.12)$$

After implementation of boundary conditions, the one-dimensional Finite Difference system analogous to (5.10) can be written explicitly as follows:

$$\begin{aligned}
&i = 1: \quad 0 + \frac{K_{\frac{1}{2}} + K_{1+\frac{1}{2}}}{\Delta x^2} \cdot H_1 - \frac{K_{1+\frac{1}{2}}}{\Delta x^2} \cdot H_2 = \frac{K_{\frac{1}{2}}}{\Delta x^2} H_o \\
&\vdots \\
&1 < i < n: -\frac{K_{i-\frac{1}{2}}}{\Delta x^2} \cdot H_{i-1} + \frac{(K_{i-\frac{1}{2}}+K_{i+\frac{1}{2}})}{\Delta x^2} \cdot H_i - \frac{K_{i+\frac{1}{2}}}{\Delta x^2} \cdot H_{i+1} = 0 \\
&\vdots \\
&i = n: -\frac{K_{n-\frac{1}{2}}}{\Delta x^2} \cdot H_{n-1} + \left\{ \frac{K_{n+\frac{1}{2}} + K_{n+\frac{1}{2}}}{\Delta x^2} - \frac{K_{n+\frac{1}{2}}}{\Delta x^2} \right\} \cdot H_i - 0 = \frac{-q_{n+\frac{1}{2}}}{\Delta x}
\end{aligned} \qquad (5.13)$$

The proposed treatment of flux boundary conditions (see the last equation above) has several advantages over other schemes proposed in the literature. First, the flux at the boundary is approximated by the same scheme as used for interior nodes (compare (5.4) to (5.11)). Second, the coefficient matrix retains the same sparsity pattern and remains symmetric after elimination of boundary values, as can be seen by inspection of (5.13). In the one dimensional case above, a tridiagonal symmetric system is obtained.

In the three-dimensional case, boundary conditions similar to (5.11) and (5.12) should be used for each node of the six planar boundaries. The resulting coefficient matrix $\underline{\underline{K}}$ is 7-diagonal symmetric, as illustrated in Figure (5.2). Each of the six off-diagonal lines contains a few zeroes corresponding to those nodes that are adjacent to one or several boundaries. For a cubic domain of size $N = n^3$, the number of such zeroes on the off-diagonal lines is only $O(n^2)$, while the size of each line is about $O(n^3)$. On the whole, only $4n^3$ matrix elements need to be defined, due to the symmetry and sparsity of the matrix. This is very small compared to the total number of elements of the matrix, $(n^3 \times n^3)$ including the zeroes. Thus, for a 1 million node grid ($n = 100$), the total number of elements in the matrix is 10^{12}, of which only 4×10^6 are actually non-zero. In addition, it is important to note that the location of the

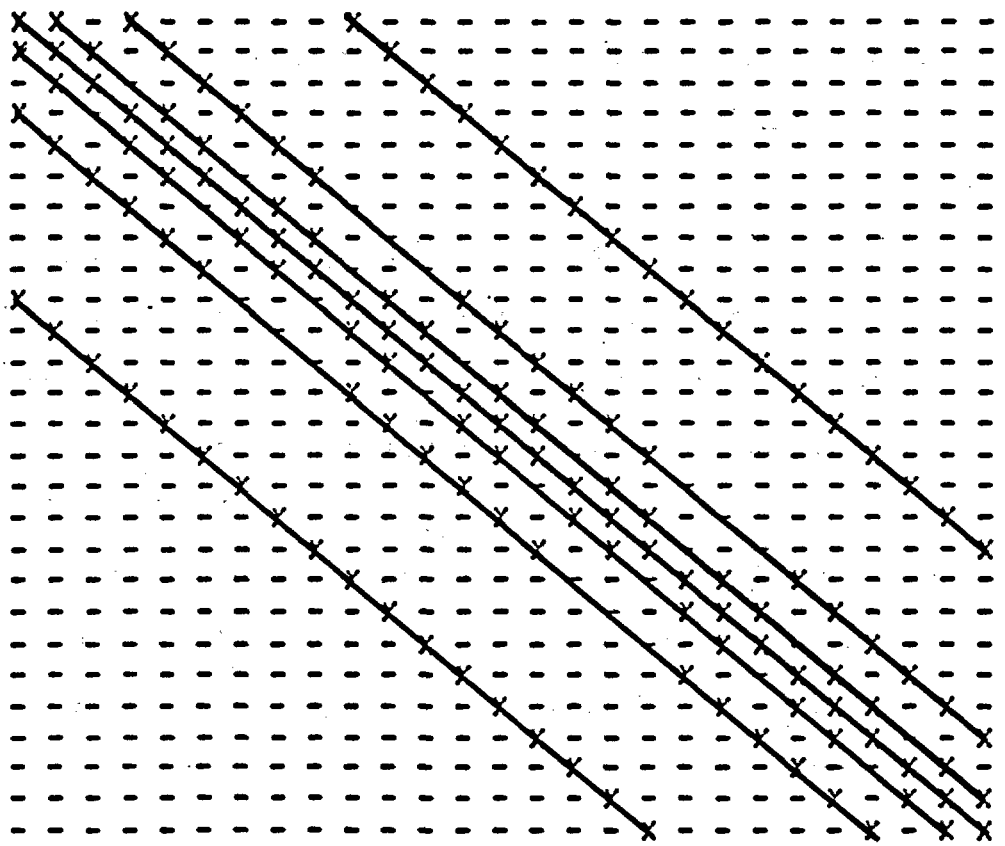


Figure 5.2 Structure of the coefficient matrix for the seven-point centered finite difference scheme, illustrated here for a 3D cubic grid with 27 internal nodes (cubic domain of side $4\Delta x$). The matrix is symmetric and has only seven non-zero diagonal lines.

non-zero elements (or non-zero lines) is known exactly. This is illustrated shown in Figure (5.2), for a small matrix of size 27 × 27 corresponding to n = 3 interior nodes along each side of the domain.

Furthermore, it can be seen by inspection that $\underline{\underline{K}}$ is symmetric positive-definite and weakly diagonal-dominant, provided that a fixed head boundary condition be specified on at least one boundary node. The matrix becomes singular in the case where all nodal boundary conditions are "fixed flux". In this case, a solution to the steady state problem exists only if the algebraic sum of in-going and out-going fluxes is identically zero; the head solution is then only defined up to an additive constant.

Finally, $\underline{\underline{K}}$ has also the "M-matrix property", that is all diagonal elements are strictly positive and all off-diagonal elements are negative or null. This property is required for certain approximate factorization methods such as Incomplete Choleski decomposition (Meijerink and Van der Vorst, 1977). More generally, many iterative solution methods rely on the system being at least symmetric positive-definite in order to ensure optimal convergence (e.g., successive overrelaxation methods). Positive-definiteness or weak diagonal dominance is also required

for the stability to round-off errors of certain matrix factorization methods, such as the Thomas algorithm for tri-diagonal matrices. This requirement is likely to extend to more general factorization methods such as the one considered in section 5.2 for the solution of the 7-diagonal finite difference system (SIP factorization).

In conclusion, the 7-point centered FD scheme seems very well suited for the application of fast iterative methods, particularly those based on approximate factorization, due to the sparsity and special algebraic structure of the coefficient matrix.

[b] - Comparison of Finite Differences with the Galerkin Method:

It may be instructive to compare the sparse FD system with the Finite Element system obtained by using the Galerkin method with tetrahedral elements and tri-linear basis functions (details can be found for instance in Huyakorn and Pinder, 1983, pp. 88, 3.5.1). One of the simplest partitions of 3D space into tetrahedral elements is obtained from a regular partition of space into hexahedral elements, each of which is further subdivided into six tetrahedra, of distinct sizes and shapes. It may be of interest to note that there exists no regular partition of the 3D space with tetrahedra all of the same size and shape;

in contrast the 2D space can be partitioned into triangles all of the same size and shape (see for instance Coxeter's book "Regular Polytopes", 1973).

The Galerkin system for the 3D saturated flow equation (5.1) takes the form:

$$\sum_{J=1}^{N} a_{IJ} \cdot H_J = b_I \qquad (I = 1, \cdots, N)$$

where $N = n^3$ is the total number of nodes for the regular orthogonal partition of space into identical hexahedra, and the matrix coefficients a_{IJ} are made up of weighted conductivities over the set of six tetrahedral elements forming an hexahedron. These coefficients take the form:

$$a_{IJ} = \sum_{H=1}^{N} \sum_{T=1}^{6} \sum_{i=1}^{3} \int_{\Omega_{H,T}} \alpha_I^{(i)} \alpha_J^{(i)} K(\underline{x}) \, d\underline{x} \qquad (5.14)$$

where tetrahedral elements were labeled "$\Omega_{H,T}$". The index H runs over all hexahedral elements (same as total number of nodes), the index T runs over the six tetrahedra comprised in one hexahedron, and the index i = 1,2,3 is related to the three independent tri-linear basis functions. It turns out that the coefficient product $\alpha_I^{(i)} \alpha_J^{(i)}$ (i = 1,2,3) is generally non-zero for 27 tetrahedra out of 6N tetrahedra on each row (equation) of the

matrix system. This yields *a priori* 27 non-zero diagonal lines in the Galerkin coefficient matrix.

In fact, most Finite Element equation solvers ignore the fine structure of the matrix, e.g. they only take into account the bandwidth of the system, (size of the band containing non-zero coefficients). The bandwidth is $2n^2 + 1$ for a cubic domain of size n^3. Taking into account the symmetry of the matrix, this implies that the number of matrix elements to be processed for solution is about n^5, compared to $4n^3$ for the 7 point FD system. In our view, these observations show that the solution of the Galerkin system would be impractical for large grids on the order of 1 million hexahedral cells ($n = 100$). Moreover, it turns out that the Galerkin system does not satisfy the "M-matrix property" mentioned above. This seems to exclude the Galerkin system as a candidate for Incomplete Choleski, and other approximate factorization methods. Note that the SIP method in particular was specifically designed for Finite Difference systems having a simple structure.

We conclude again that the 7-point centered Finite Difference scheme appears the most suited for the solution of very large flow problems with fluctuating conductivities, due to

the special algebraic properties and sparsity of the FD system. In comparison, "higher order" discretization methods such as Galerkin lead to more complex matrix structures, particularly in the three-dimensional case. We feel that the advantage of using higher order or smoother numerical approximations may well be offset by the significant increase in computational work for three-dimensional, high resolution simulations.

5.1.3: Finite difference in space-time for transient unsaturated flow

We now extend our finite difference discretization method to the case of unsaturated flow, or more generally "partially saturated flow", in a statistically heterogeneous porous medium. In what follows, we focus particularly on the case of transient flow, as this is of interest for applications like local infiltration in unsaturated soils. However, the cases of steady unsaturated flow, mixed saturated/unsaturated flow, as well as the case of purely saturated flow studied just above, are all embodied in the general unsaturated flow problem treated in this section.

[a] Governing Equation and Constitutive Soil Properties:

The general unsaturated flow problem in a heterogeneous porous medium is governed by the continuity equation:

$$\frac{\partial}{\partial t}[S(h,\underline{x}) + \theta(h,\underline{x})] = -\frac{\partial q_i}{\partial x_i} \quad (i = 1,2,3) \qquad (5.15)$$

and the generalized Darcy equation:

$$q_i = -K(h,\underline{x}) \cdot \frac{\partial}{\partial x_i}(h + g_j x_j) \qquad (5.16)$$

where:

- $S(h,\underline{x})$ is the storage term due solely to water-soil compressibility under positive pressures (cm^3/cm^3)
- $\theta(h,\underline{x})$ is the pressure-dependent, spatially variable volumetric soil water content relative to an incompressible soil matrix (cm^3/cm^3)
- q_i is the flux vector, or specific discharge rate (cm/s)
- $K(h,\underline{x})$ is the pressure-dependent, spatially variable unsaturated hydraulic conductivity (cm/s)
- h is the water pressure head relative to atmospheric pressure (cm)
- g_i is a cosine vector, corresponding to the unit acceleration of gravity taken with a minus sign.

The generalized Darcy equation (5.16) calls for explanations. First, note that the "gravity vector" g_i has components (-1,0,0) if the x_1 axis is taken to be vertical downwards. More generally, $g_i = (\sin\gamma, 0, \cos\gamma)$ if there is an angle γ between x_3 and the upward vertical axis. This formulation makes it possible to simulate infiltration onto hill slopes and similar unsaturated flow problems involving sloping faces. Second, it should be observed that the hydraulic head H appears implicitly in the generalized Darcy equation (5.16), in the form:

$$H = h + g_j x_j \qquad (j = 1,2,3)$$

i.e., as the sum of a pressure potential and a gravity potential. Note that the "pressure head" h may attain very large negative values in a dry soil, especially in clay soils or in the presence of active plant roots. In these cases, h should be interpreted as a thermodynamic potential. The minus hydraulic head (-H) stands for the energy that must be produced in order to bring soil water to its free state at a plane of reference such as soil surface ($g_i x_j = 0$). Soil water is in its free state when $H = 0$, has negative energy when $H < 0$, and positive energy when $H > 0$.

We now introduce specific constitutive relations for the unsaturated conductivity $K(h,\underline{x})$, the soil moisture retention

curve $\theta(h,\underline{x})$, and the elastic storage term $S(h,\underline{x})$. First of all, we assume that the elastic storage term plays no role for purely unsaturated flow, provided that positive pressures never appear within the flow domain. This may be realistic in the case of infiltration at low flow rate. On the other hand, it should be recognized that perched water tables or locally saturated zones are likely to appear in the case of high rate infiltration in a heterogeneous medium. For this reason, the elastic storage term was retained in the numerical code, assuming a simple dependence on water pressure as follows:

$$S(h,x) = \begin{cases} S_s(\underline{x}) \cdot h & \text{if } h \geq 0 \\ 0 & \text{if } h \leq 0 \end{cases} \qquad (5.18)$$

where $S_s(\underline{x})$ is the specific storativity (cm^{-1}) which accounts for the compressibility of water and the solid porous matrix under positive pressures. For simplicity, S_s can be taken constant when the flow system is mostly in the unsaturated regime. In this work, S_s was neglected altogether since most unsaturated flow simulations concerned the case of low rate infiltration in dry soils (Section 5.4.3 and Chapter 7).

The soil moisture retention curve $\theta(h,\underline{x})$ was assumed to take the form of a multi-parameter nonlinear function of pressure h, possibly with spatially variable parameters. The particular

model we used for unsaturated flow simulations is the well-documented Van Genuchten model (Van Genuchten, 1978 and 1980). With spatially variable parameters, this model can be expressed as:

$$\theta(h,\underline{x}) = \theta_r(\underline{x}) + \frac{\theta_s(\underline{x}) - \theta_r(\underline{x})}{\{1 + (-\beta h)^n\}^{1-1/n}} \qquad (5.19)$$

where β is a scale factor (cm^{-1}), n is a real dimensionless parameter (identified as "VGN" in the flow simulator), θ_s is the saturated soil water content, and θ_r is the residual water content at very high negative pressures. Note that the term $(\theta_s(\underline{x}) - \theta_r(\underline{x}))$ can be thought of as a spatially variable "effective porosity". The β and n parameters could also be taken spatially variable. However the results of the linearized spectral theory suggest that the effects of the spatial variability of $\theta(h,\underline{x})$ are small compared to those due to $K(h,\underline{x})$: see Mantoglou 1984, and Mantoglou and Gelhar, 1987. For this reason we will assume constant parameters in the water retention curve for stochastic unsaturated flow simulations (Chapter 7), although some of the preliminary numerical experiments developed in the present chapter will include uniformly layered soil systems where both the $\theta(h)$ and $K(h)$ curves vary from layer to layer (section 5.4.3).

The most relevant feature of the constitutive relation

(5.19) is its nonlinear (logistic curve) shape. This is illustrated in Figures (5.3) and (5.4) for two different types of soils (from Ababou, 1981). The location of the inflexion point is of particular interest, since this is the point where the specific soil moisture capacity $C(h) = \frac{\partial \theta}{\partial h}$ attains its maximum. This inflexion point is given below in close form:

$$\begin{cases} h_{max} = -\frac{1}{\beta} \cdot (1 - \frac{1}{n})^{1/n} \\ C_{max} = \beta \cdot (\theta_s - \theta_r) \cdot (n - 1) \cdot \frac{(m)^m}{(m + 1)^{m+1}} \end{cases} \qquad (5.20)$$

where $m = 1-1/n$. For coarse or sandy soils the maximum capacity occurs at relatively high pressures (h_{max}) with a narrow peak compared to finer soils. For example, the maximum capacities for the Dek Sand and Montfavet Silt soils depicted in Figure (5.3) and (5.4) were, respectively:

- Sand: $C_{max} = 2.82 \times 10^{-3}\ cm^{-1}$, $h_{max} = -24.5$ cm
- Silt: $C_{max} = 0.33 \times 10^{-3}\ cm^{-1}$, $h_{max} = -344.$cm

Furthermore, the shape of the $\theta(h)$ curve indicates that the soil moisture capacity must be approximately constant in the pressure range:

$$|h-h_{max}| \ll \beta^{-1} \qquad (5.21)$$

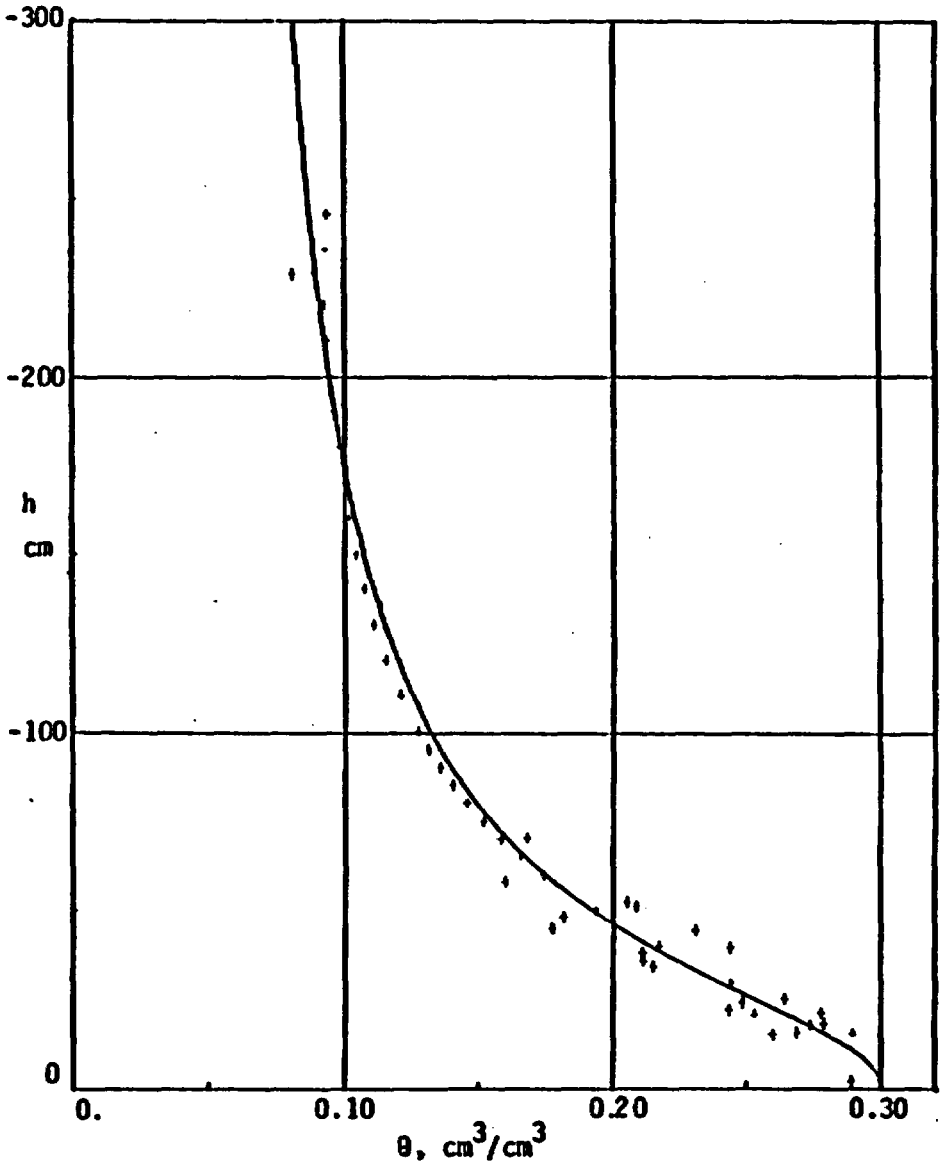


Figure 5.3(a) Soil water retention curve θ(h) for the "Dek sand" of Senegal. The solid line represents the Van Genuchten function fitted to data points (from Ababou, 1981).

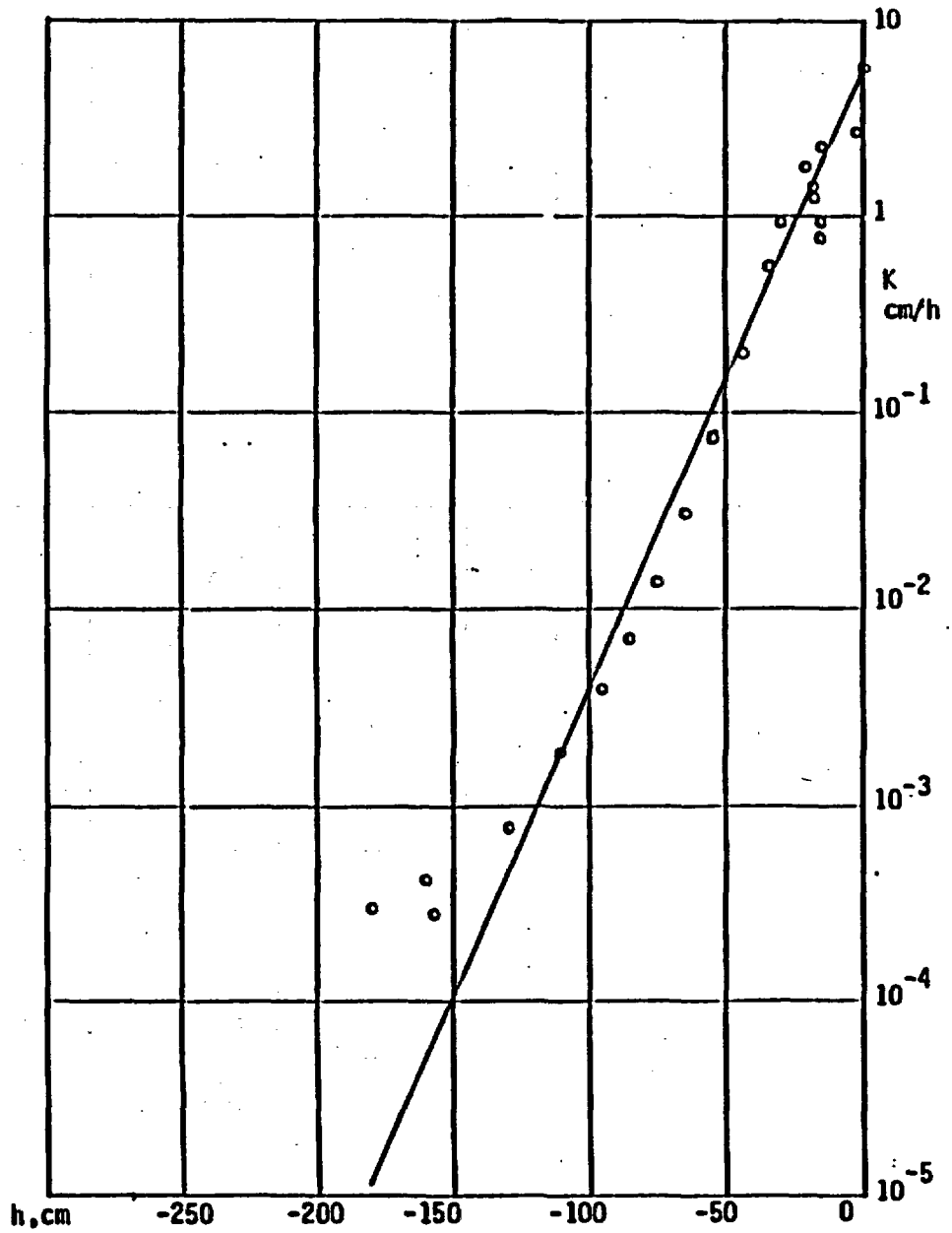


Figure 5.3 (b) Unsaturated conductivity curve K(h) for the "Dek sand" of Senegal. The solid line represents the exponential conductivity curve fitted to data points (from Ababou, 1981).

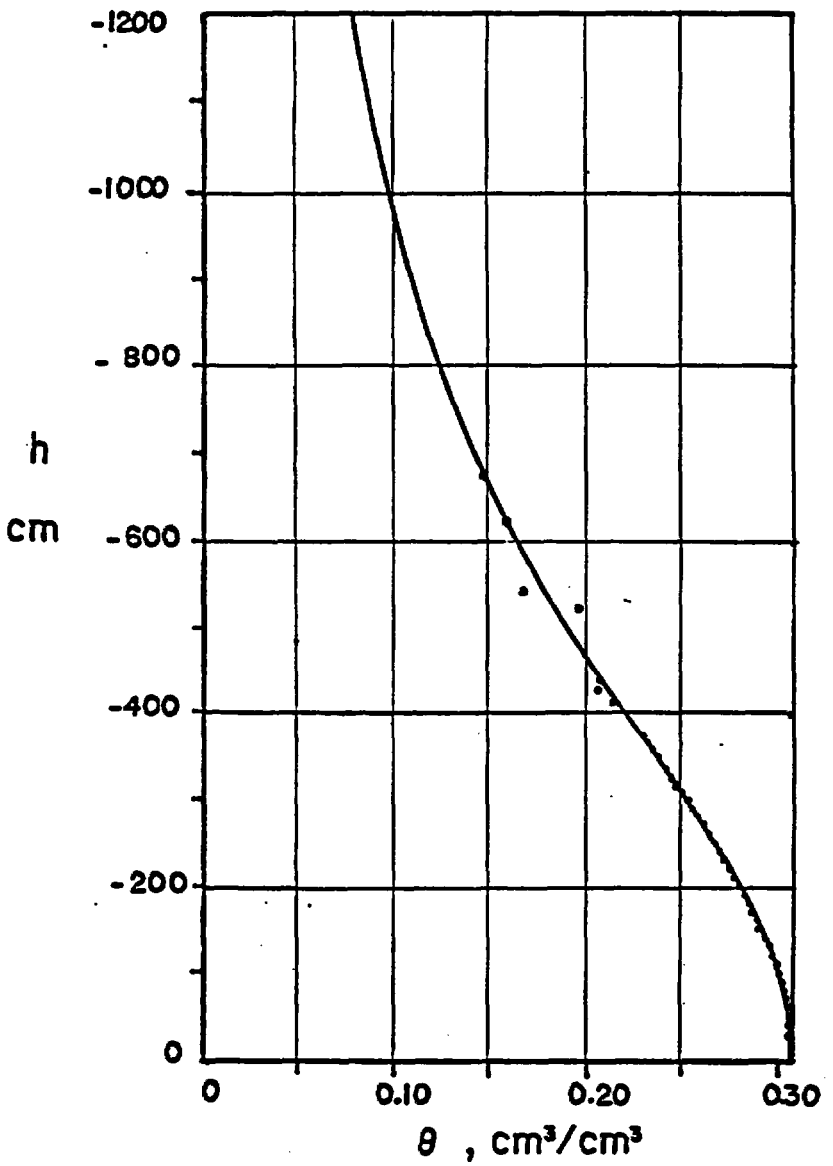


Figure 5.4 (a) Soil water retention curve θ(h) for the Montfavet silt, a loess soil from the south of France. The Van Genuchten curve (solid line) was fitted to data points (Ababou, 1981).

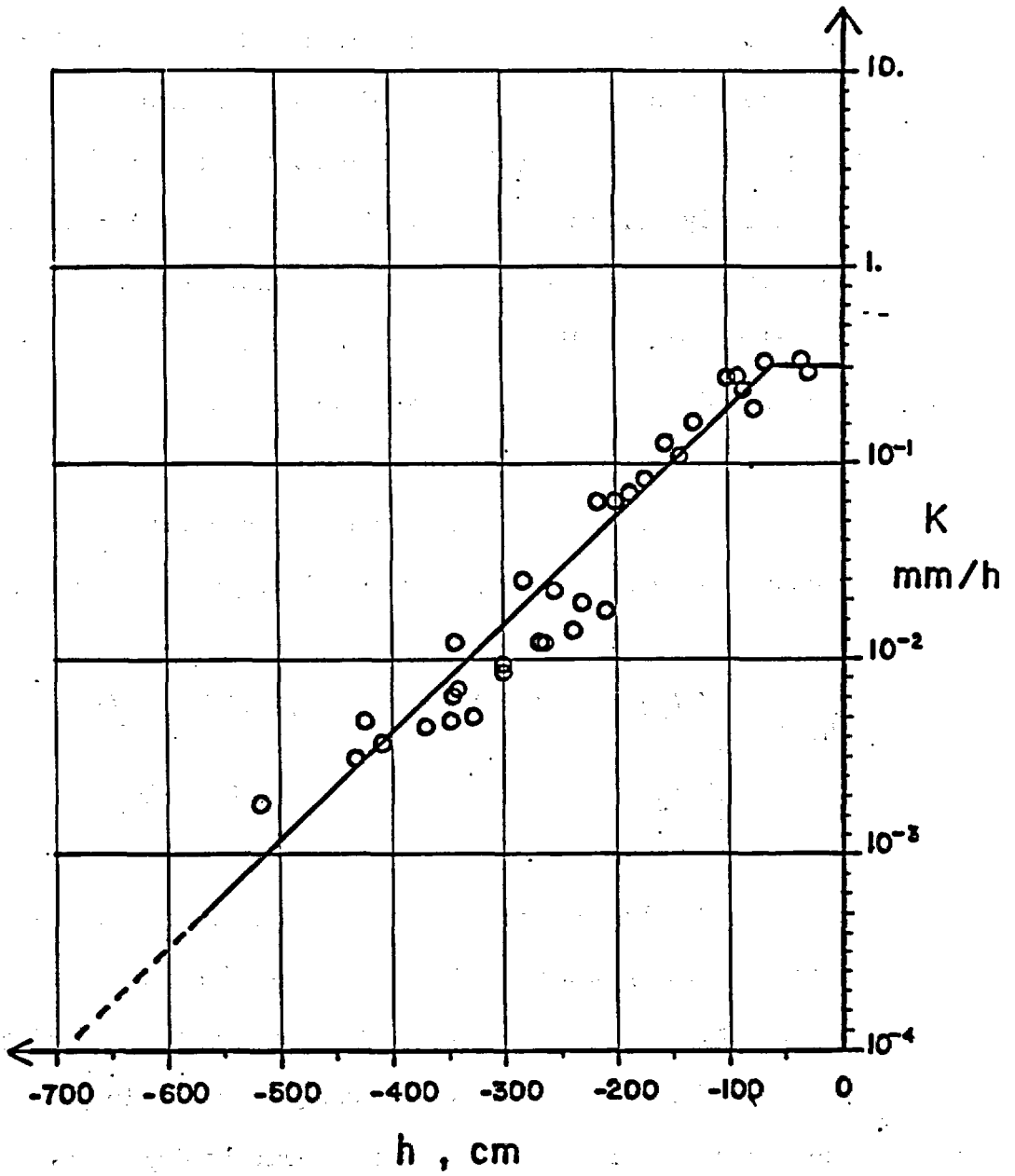


Figure 5.4 (b) Unsaturated conductivity curve K(h) for the Montfavet silt. The exponential conductivity curve (solid line) was fitted to data points (Ababou, 1981).

In the examples above, β was about 2.3×10^{-2} for the sand, and 2.3×10^{-3} for the silt. Accordingly, it appears that the soil moisture capacity is approximately constant and close to C_{max} in the pressure ranges (-90cm to -10cm) and (-650cm to -150cm) respectively for the sand and silt soils. Within these ranges, the simplifying assumption of a constant capacity used by Mantoglou and Gelhar (1987) in their spectral theory of unsaturated flow appears to hold approximately.

Finally, we chose an exponential model for the unsaturated conductivity-pressure relation; as shown below:

$$K(h,\underline{x}) = \begin{cases} K_s(\underline{x}) \cdot \exp\{\alpha(\underline{x})\cdot(h - h_b(\underline{x}))\} & \text{if } h \leq h_b(x) \\ K_s(\underline{x}) & \text{if } h \geq h_b(x) \end{cases} \qquad (5.22)$$

When the "bubbling pressure head" h_b is taken to be null, equation (5.22) becomes identical to the model used in the spectral theory of Mantoglou and Gelhar. Figures (5.3) and (5.4) show that the exponential conductivity model is in good agreement with measured values for wet and moderately dry soils. However, these and other data also suggest that the exponential rate of decrease of K(h) is not sustained as the soil dries up below a certain pressure (e.g., $h \simeq -600$ cm for the silt). This limitation of the exponential model will be taken into account in

our simulations of field infiltration problems, by using the field measured initial conductivity (rather than the field-measured pressure) for an estimate of the initial pressure based on the exponential conductivity model. The initial pressure obtains by plugging $K = K_{in}$ (field-measured) into the inverse relation:

$$h = \frac{1}{\alpha} \ell n(\frac{K}{K_s})$$

where α and K_s represent using mean values, in lieu of $\alpha(\underline{x})$ and $K_s(\underline{x})$. This expedient procedure may be useful in cases where the field-measured initial pressure is outside the range of validity of the exponential model (very low pressure) whereas most of the unsaturated flow process occurs at higher pressures well within the range of validity of the exponential model.

In view of the results obtained by the spectral perturbation method of Mantoglou and Gelhar, our main focus will be the study of infiltration problem with the exponential conductivity curve having both $K_s(\underline{x})$ and $\alpha(\underline{x})$ random fields in 3D space. This, combined with the strong nonlinearity of $K(h,\underline{x})$ with respect to h, makes the numerical solution of the unsaturated flow problem a difficult task indeed.

[b] Time Discretization:

The time discretization of the transient unsaturated flow equation is now being developed. For convenience, let us merge the elastic storage and the soil moisture terms into a single term:

$$\theta(h,\underline{x}) := S(h,\underline{x}) + \theta(h,\underline{x})$$

Plugging the Darcy equation into the continuity equation, one obtains the governing nonlinear flow equation for pressure head or "Richards equation":

$$\frac{\partial\theta(h,\underline{x})}{\partial t} = \frac{\partial}{\partial x_i}\left[K(h,\underline{x}) \cdot \left(\frac{\partial h}{\partial x_i} + g_i\right)\right] = - L(h,\underline{x}) \quad (5.23)$$

For the time integration scheme, we choose the fully implicit (Backward Euler) two-point finite difference scheme. Denoting $(-L(h,\underline{x}))$ the spatial operator on the right-hand side of (5.23), the fully implicit time discretization scheme can be expressed as the first order finite difference approximation:

$$\frac{\theta^{n+1}(h,x) - \theta^{n}(h,x)}{\Delta t_{n+1}} \simeq - L^{n+1}(h,\underline{x}) \quad (5.24)$$

where $\Delta t_{n+1} = t_{n+1} - t_n$. Alternatively, this finite difference approximation could also be interpreted as a time integral approximation. Indeed, by integrating the exact equation (5.23) between times (t_n, t_{n+1}) one obtains:

$$\theta^{n+1}(h,\underline{x}) - \theta^{n}(h,\underline{x}) = - \int_{t_n}^{t_{n+1}} L(h,\underline{x})\, dt$$

and, using the mean value theorem:

$$\frac{\theta^{n+1}(h,x) - \theta^{n}(h,x)}{\Delta t_{n+1}} = - \{\tau \cdot L^{n+1}(h,\underline{x}) + (1 - \tau) \cdot L^{n}(h,\underline{x})\} \quad (5.25)$$

where $0 \leq \tau \leq 1$. In particular for $\tau = 1$ one obtains the fully implicit finite difference approximation (5.24). The more general class of implicit scheme corresponds to $\frac{1}{2} \leq \tau \leq 1$. The Crank-Nicholson scheme in particular corresponds to the case $\tau = 1/2$.

Briefly, our particular choice $\tau = 1$ was based on results of stability theory and various numerical experiments in the literature. First of all, it is well known that implicit schemes are unconditionally stable, whereas explicit schemes require for stability a stringent constraint on the time step. For the simple heat equation $\frac{\partial u}{\partial t} = D \cdot \frac{\partial^2 u}{\partial x_i \partial x_i}$, the stability constraint takes the form (Ames, 1977):

$$\Delta t \leq \left[2D \cdot \sum_i \frac{1}{\Delta x_i^2}\right]^{-1}.$$

This observation holds also for more complex cases such as the nonlinear unsaturated flow equation. Of course, the exact form of the stability constraint depends in fact on the spatial discretization scheme and the linearization scheme. In any case the resulting stability constraint is qualitatively of the same form as shown above. In particular, note that the soil moisture diffusivity $D = K/C$ may become quite large in a wet zone, so that the time step may have to be dramatically small in order to satisfy the stability constraint. Taking the Dek sand of Figures (5.3)-(5.4) as an example, and using a value of pressure corresponding to the maximum soil moisture capacity, we find that $\Delta t \leq 4.3$ sec is needed to ensure the stability of the explicit scheme on a 3D grid with $\Delta x_i = 5$ cm. For "wetter" conditions, the stability constraint would be even more drastic. This justifies the use of the unconditionally stable implicit scheme for time integration.

Our second remark is about the choice of the fully implicit scheme ($\tau = 1$) in preference to other implicit schemes such as Crank-Nicholson ($\tau = 1/2$). Numerical experiments tend to show that the fully implicit scheme is particularly efficient for the case of the *nonlinear* flow equation at hand (Vauclin et al., 1979), although there is no theoretical evidence in favor of one

type of implicit scheme over the other for strongly *nonlinear* equations.

[c] Linearization and Spatial Discretization:

We now focus on the nonlinear semi-discretized equation (5.24). In order to obtain a tractable *linear* system, we propose an approximate linearization of (5.24) by way of iterative corrections based on a modified Picard iteration scheme. The procedure is best explained in two steps: (i) linearization of the right hand side spatial operator, and (ii) linearization of the left hand side temporal operator. This is described in detail below:

(i) - We use a Picard iteration scheme to approximate the nonlinear equation (5.24) into a sequence of equations ($k = 0,1,2 \cdots$) where the conductivity appears linearly as follows:

$$\frac{\theta^{n+1,k+1}(h,x) - \theta^{n}(h,x)}{\Delta t_{n+1}} \simeq \frac{\partial}{\partial x_i}\left[K^{n+1,k}(h,\underline{x}) \cdot \left[\frac{\partial h^{n+1,k+1}}{\partial x_i} + gi\right]\right]$$

Note that the conductivity on the right-hand side is evaluated from the previous iteration level. By substituting to both sides the "residual":

$$R^{n+1,k} = - \left\{ \frac{\theta^{n+1,k} - \theta^{n}}{\Delta t_{n+1}} - \frac{\partial}{\partial x_i} \left[K^{n+1,k} \cdot \left[\frac{\partial h^{n+1,k}}{\partial x_i} + g_i \right] \right] \right\}$$

we obtain equivalently the "modified" Picard scheme:

$$\frac{\theta^{n+1,k+1} - \theta^{n+1,k}}{\Delta t_{n+1}} - \frac{\partial}{\partial x_i} \left[K^{n+1,k} \cdot \frac{\partial}{\partial x_i} \left(h^{n+1,k+1} - h^{n+1,k} \right) \right] \simeq R^{n+1,k} \tag{5.26}$$

which is computationally more stable with respect to round-off errors. For clarity, note that n represents the time level, while k represents the iteration level. At this point, we have only linearized the spatial operator $\frac{\partial}{\partial x_i} [K(h) \frac{\partial}{\partial x_i}(h+g_i)]$. However the discrete storage term involving $\theta(h)$ is still nonlinear.

(ii) - In order to obtain a fully linear equation, the storage term is now linearized by applying the mean value theorem as follows:

$$\theta^{k+1} - \theta^{k} = \int_{h^k}^{h^{k+1}} C(h)\, dh$$
$$= (h^{k+1} - h^{k}) \cdot C[(1 - \tau)h^{k} + \tau h^{k+1}]$$

where τ is some number in the interval $[0,1]$. The choice $\tau = 0$ would lead to a linear equation in h, as required. We prefer a more stable approximation similar to the chord-slope

approximations proposed by Huyakorn, et al. (1984) and Milly (1985), that is:

$$\begin{cases} \dfrac{\theta^{k+1} - \theta^{k}}{\Delta t_{n+1}} \simeq \dfrac{\hat{C}(h^{k})}{\Delta t_{n+1}} \cdot (h^{k+1} - h^{k}) \\ \\ \hat{C}\,(h^{k}) \quad = \dfrac{\theta(h^{k}) - \theta(h^{o})}{h^{k} - h^{o}} \end{cases} \tag{5.27}$$

where all variables are for time level (n+1), except h^{o} which stands for "iteration level (0)". In other words, h^{o} is the known solution at the previous time level n.

Combining (5.26) and (5.27) we finally obtain a fully linear semi-discretized equation in terms of the incremental pressure head δh between successive iterations:

$$\boxed{\frac{\hat{C}^{n+1,k}}{\Delta t_{n+1}} \cdot \delta h - \frac{\partial}{\partial x_i}\left[K^{n+1,k} \cdot \frac{\partial}{\partial x_i}(\delta h)\right] \simeq R^{n+1,k}} \tag{5.28}$$

where:

$$\delta h \quad = h^{n+1,k+1} - h^{n+1,k}.$$

and:

$$R^{n+1,k} = -\left\{\frac{\hat{C}^{n+1,k}}{\Delta t_{n+1}}(h^{n+1,k} - h^{n}) - \frac{\partial}{\partial x_i}\left[K^{n+1,k}\left[\frac{\partial h^{n+1,k}}{\partial x_i} + g_i\right]\right]\right\}$$

In summary, equation (5.28) is a linearized, semi-discrete

approximation of the original unsaturated flow equation (5.23), which can be used as a starting point for further discretization in space. This is examined next.

The spatial operators appearing on the left and right-hand sides of (5.28) are now discretized by using the 7-point centered finite difference scheme, in the same fashion as in the previous section dealing with saturated flow (Eqs. 5.4-5.7). For convenience, the two spatial operators appearing in (5.28) will be designated as:

$$L_K(Y) = - \frac{\partial}{\partial x_i} (K \frac{\partial Y}{\partial x_i})$$

$$L_G(Y) = - g_i \cdot \frac{\partial K}{\partial x_i}.$$

Applying the 7-point finite difference scheme to the L_K-operator leads to an expression similar to that obtained for saturated flow (Eqs. 5.1 and 5.6), with the mid-nodal unsaturated conductivities defined as:

$$K_{i+\frac{1}{2},j,k} = K(h_{i+\frac{1}{2},j,k}, x_{i+\frac{1}{2},j,k}).$$

Furthermore, we use again the geometric mean weighting scheme to evaluate the mid-nodal conductivities:

$$\hat{K}_{i+\frac{1}{2},j,k} = \sqrt{K(h_{ijk},x_{ijk}) \cdot K(h_{i+1,j,k},x_{i+1,j,k})}$$

In the particular case of the exponential conductivity model (5.22) with a zero bubbling pressure, this gives:

$$\hat{K}_{i+\frac{1}{2},\ j,k} = \sqrt{K_s(x_{i+1,j,k}) \cdot K_s(x_{i,j,k})} \cdot \exp\left\{\frac{\alpha(x_{i+1,j,k}) \cdot h_{i+1,j,k} + \alpha(x_{i,j,k}) \cdot h_{i,j,k})}{2}\right\} \tag{5.29}$$

This particular choice of conductivity weighting was guided by the results of the spectral theory concerning the effective conductivity in a random unsaturated soil (see Mantoglou and Gelhar, 1987). According to these authors, the effective unsaturated conductivity is of the form:

$$K_{eff} \sim K_G \cdot \exp\{\langle \alpha h \rangle\}$$

where K_G is the ensemble geometric mean of the random saturated conductivity. The conductivity weighting scheme (5.29) may be justified by analogy with this result.

The L_G-operator may be approximated in the same fashion as the L_K-operator. Using the same index notation as before (see equation 5.6) we obtain the following centered finite difference approximation:

$$\hat{L}_G(Y) = -\left\{\frac{g_1}{\Delta x_1}(K[i+\tfrac{1}{2}] - K[i-\tfrac{1}{2}]) + \frac{g_2}{\Delta x_2}(K[j+\tfrac{1}{2}] - K[j-\tfrac{1}{2}]) + \frac{g_3}{\Delta x_3}(K[k+\tfrac{1}{2}] - K[k-\tfrac{1}{2}])\right\} \qquad (5.30)$$

Mid-nodal conductivities like $K[i+\tfrac{1}{2}] = K_{i+\frac{1}{2},j,k}$ are as defined just above.

The fully discretized, linearized unsaturated flow system obtains by plugging $\hat{L}_G$ of (5.30) and $\hat{L}_K$ of (5.6) into equation (5.28). The resulting finite difference system takes the form:

$$\boxed{\left(\frac{C}{\Delta t} \cdot \underline{\underline{I}} + \underline{\underline{K}}\right)^{n+1,k} \cdot \underline{\delta h} = \underline{r}^{n+1,k}} \qquad (5.31)$$

where $\underline{\underline{I}}$ is the identity matrix, $\underline{\underline{K}}$ is the unsaturated conductivity matrix, δh is the increment of pressure between two iteration levels (k,k+1), and $\underline{r}$ is the vector of residuals including also the vector of boundary conditions. It is worth noting that the conductivity matrix $\underline{\underline{K}}$ has exactly the same form as for the saturated flow equation, with the unsaturated mid-nodal conductivities of (5.29) replacing the saturated mid-nodal conductivities of (5.28). As a consequence, the coefficient matrix:

$$\underline{\underline{A}} = \frac{C}{\Delta t} \cdot \underline{\underline{I}} + \underline{\underline{K}}$$

has the same sparsity pattern for saturated and unsaturated flow. In both cases, the matrix is a 7-diagonal symmetric as shown in Figure (5.1).

Furthermore, the coefficient matrix is strictly diagonal- dominant in the case where the storage term ($C/\Delta t$) is strictly positive for all times and all locations, as would occur for transient infiltration in a moderately dry soil. As a consequence, the iterative matrix solver could converge much faster in the transient case than in the steady state case (due to better matrix condition). This suggests that there will be a trade-off between the requirement of strong diagonal dominance for faster solution (small Δt), and the need to minimize the number of time steps (large Δt). In addition, a small time step may be required for fast convergence of the nonlinear-Picard iterations. The numerical experiments of section 5.4 will help determine the appropriate strategy in that respect.

5.2: Statistical Truncation Error Analysis for Linear Random Flow Problems

In this section, we focus on the evaluation of numerical errors due to the discretization of the stochastic flow equation. In particular, we develop in detail a *statistical truncation error analysis* of the finite difference approximation of the steady state saturated flow equation with random field conductivities. This will lead to some useful results concerning the order of accuracy of the finite difference method for a certain class of stochastic equations (random heat equations). Both the method and results appear to be new in view of the current literature on numerical analysis. The final results will be summarized and discussed at the end of this section, notably in terms of numerical requirements like grid resolution. For completeness, note that the nonlinear problem of transient unsaturated flow will be analyzed separately in a forthcoming section, however in a much more qualitative way.

Because the statistical truncation error analysis of this section is somewhat intricated, we feel that it may be useful to outline here the main features of our approach. Our purpose is to obtain a closed form statistical evaluation of the finite difference error, that is the error on the finite difference solution due solely to truncation errors, assuming

that an exact solution of the finite difference system can be achieved. The method we use is applicable to linear problems, such as the linear system arising from the stochastic saturated flow equation. Briefly, we begin by evaluating analytically the truncation error of the finite difference operator. We then obtain an equivalent partial differential equation that governs the error on the hydraulic head. This equation has the same structure as the governing equation, except that it is driven by the truncation error (forcing function). To solve this "error equation", we use a first order spectral perturbation method with the usual assumptions of stationarity and ergodicity (the "exact" solution is also evaluated by this method). This gives finally the error in the form of a random field with known spectral density. The root-mean-square error on the hydraulic head obtains by computing the variance of the random field error, and taking the square root. A similar procedure is then used to evaluate the error in the flux vector. The reader not interested in the details may jump to the "summary and discussion" given in subsection 5.2.4.

5.2.1 Governing equation for the numerical head error

Recall that the exact equation governing the head variable is of the form $L(H) = 0$, where $L(H)$ is the partial differential operator:

$$L(H) = \frac{\partial}{\partial x_m}\left(K(x)\frac{\partial H}{\partial x_m}\right) \qquad (m = 1,2,3)$$

The FD solution $\hat{H}$ defined on a grid (I_1, I_2, I_3) satisfies the FD equations $\hat{L}(\hat{H}) = 0$, where $L(\hat{H})$ is the difference operator:

$$\hat{L}(\hat{H}) = \sum_{m=1,2,3} \frac{1}{(\Delta x_m)^2}\left\{\hat{K}_{I_{m+\frac{1}{2}}} \cdot \hat{H}_{I_{m+1}} - \left[\hat{K}_{I_{m+\frac{1}{2}}} + K_{I_{m-\frac{1}{2}}}\right] \cdot \hat{H}_o + \hat{K}_{I_{m-\frac{1}{2}}} \cdot \hat{H}_{I_{m-1}}\right\}$$

Note that m indicates the three directions x_m, and we used the shorthand notations:

$$\hat{H}_o \rightarrow \hat{H}(I_1, I_2\ I_3)$$

$$\hat{H}_{I_1+1} \rightarrow \hat{H}(I_1+1, I_2, I_3)$$

$$\hat{K}_{I_1+\frac{1}{2}} \rightarrow \hat{K}(I_1+\frac{1}{2},\ I_2, I_3)$$

(etc.)

The mid-nodal conductivities were evaluated by the geometric mean weighting scheme:

$$\hat{K}_{I_1+\frac{1}{2}} = \sqrt{K(I_1,I_2,I_3) \cdot K(I_1+1,I_2,I_3)}$$

(etc.).

and the flux vector:

$$q_m = - K \frac{\delta H}{\delta x_m}$$

was evaluated at mid-nodal points by:

$$\hat{q}_{I_m+\frac{1}{2}} = - \hat{K}_{I_m+\frac{1}{2}} \cdot \frac{\hat{H}_{I_m+1} - H_0}{\Delta x_m}.$$

Finally, we now prescribe *Dirichlet boundary conditions* on all boundaries, since these conditions can be expressed exactly in the finite difference formulation. The introduction of Neuman conditions would complicate unnecessarily the forthcoming error analysis.

The *truncation error* is defined as the difference between the exact and approximate operators (both operating on the exact solution H):

$$T(H) = \hat{L}(H) - L(H)$$

where H is evaluated at the grid points $\underline{I} = (I_1, I_2, I_3)$. *Once computed as a function of* H, *the truncation error could also be evaluated as a function of the approximate FD solution* $\hat{H}_I$, or more precisely as a function of an equivalent continuous FD solution $\tilde{H}$ which takes the values of the discrete FD solution $\hat{H}_i$ at the grid points. Thus we can write:

$$T(\tilde{H}) = \hat{L}(\tilde{H}) - L(\tilde{H})$$

where $\tilde{H}(\underline{x})$ is a continuous function such that $\tilde{H}(x_I) = \hat{H}_I$. Observe that the term $\hat{L}(\tilde{H})$ can be eliminated since it is identically zero at the grid points (by definition of $\hat{H}$). This yields immediately a partial differential equation governing the "Equivalent Continuous Finite Difference Solution":

$$\frac{\partial}{\partial x_m}\left[K(\underline{x})\,\frac{\partial \tilde{H}}{\partial x_m}\right] + T(\tilde{H}) = 0. \qquad (5.32)$$

Similarly, the continuous solution error can be defined as:

$$\delta\tilde{H}(\underline{x}) = \tilde{H}(\underline{x}) - H(\underline{x})$$

$$\delta\tilde{H}_I = \hat{H}_I - H(x_I). \qquad (5.33)$$

Ultimately, one would like to obtain an estimate for the error

$\delta\tilde{H}_I$ at the grid points. This can be achieved by developing $\delta H(\underline{x})$ into a power series of (Δx_m):

$$\delta\tilde{H}_I(\underline{x}) = \delta\tilde{H}^{(0)} + \Delta x_m \cdot \delta\tilde{H}_m^{(1)} + (\Delta x_m)^2 \cdot \delta\tilde{H}_m^{(2)} + \cdots \qquad (5.34)$$

Obviously, the zero-order term $\delta H^{(0)}$ should vanish if the FD approximation is consistent. In fact it will turn out that the zero-order and all odd-order terms in the expansion vanish.

To see that $\delta\tilde{H}^{(0)}$ vanishes, we plug the identity $\tilde{H} = H+\delta\tilde{H}$ into (5.32) to obtain:

$$\boxed{\begin{array}{l} \dfrac{\partial}{\partial x_m}\left[K(\underline{x}) \cdot \dfrac{\partial(\delta\tilde{H})}{\partial x_m} \right] + T(H) + T(\delta\tilde{H}) = 0 \\ \\ \delta H = 0 \quad \text{on all boundaries (exact BC's)} \end{array}} \qquad (5.35)$$

where we used the fact that T is a linear function, due to the linearity of the flow equation. Combining (5.34) and (5.35), it appears that $\delta\tilde{H}^{(0)}$ vanishes if $\lim_{\Delta x \to 0} T(H) = 0$, that is, provided that the finite difference scheme is a consistent approximation. We will see shortly that this is indeed the case.

We now proceed to evaluate explicitly the truncation error function T(H) in order to obtain the equation governing the

solution error δH (from (5.34) and (5.35)). This is done by using Taylor series expansions such as:

$$H_{i+1}-H_{i+\frac{1}{2}} = \frac{1}{2}\left(\frac{\partial H}{\partial x}\right)_{i+\frac{1}{2}} \cdot \Delta x + \frac{1}{8} \cdot \left(\frac{\partial^2 H}{\partial x^2}\right)_{i+\frac{1}{2}} \cdot \Delta x^2 + \frac{1}{48}\left(\frac{\partial^3 H}{\partial x^3}\right)_{i+\frac{1}{2}} \Delta x^3 + \cdots$$

$$H_{i+\frac{1}{2}}-H_{i} = \frac{1}{2}\left(\frac{\partial H}{\partial x}\right)_{i+\frac{1}{2}} \Delta x - \frac{1}{8}\left(\frac{\partial^2 H}{\partial x^2}\right)_{i+\frac{1}{2}} + \frac{1}{48}\left(\frac{\partial^3 H}{\partial x^3}\right)_{i+\frac{1}{2}} \Delta x^3 + \cdots$$

whence:

$$\frac{H_{i+1}- H_{i}}{\Delta x} = \left(\frac{\partial H}{\partial x}\right)_{i+\frac{1}{2}} + \frac{\Delta x^2}{24} \cdot \left(\frac{\partial^3 H}{\partial x^3}\right)_{i+\frac{1}{2}} + O(\Delta x^4) \qquad (5.36)$$

$$\frac{Y_{i+\frac{1}{2}} -Y_{i-\frac{1}{2}}}{\Delta x} = \left(\frac{\partial Y}{\partial x}\right)_{i} + \frac{\Delta x^2}{24}\left(\frac{\partial^3 Y}{\partial x^3}\right)_{i} + O(\Delta x^4). \qquad (5.37)$$

Plugging $Y = K\frac{\partial H}{\partial x}$ into the last equation yields the flux divergence:

$$\frac{1}{\Delta x}\left[\left(K_{i} \cdot \left(\frac{\partial H}{\partial x}\right)_{i+\frac{1}{2}} - K_{i+\frac{1}{2}} \cdot \left(\frac{\partial H}{\partial x}\right)_{i-\frac{1}{2}}\right] =$$

$$\left[\frac{\partial}{\partial x}\left(K\frac{\partial H}{\partial x}\right)\right] + \frac{\Delta x^2}{24} \cdot \left[\frac{\partial^3}{\partial x^3}\left(K\frac{\partial H}{\partial x}\right)\right]_{i} + O(\Delta x^4). \qquad (5.38a)$$

Substituting (5.36) for the terms $\left(\frac{\partial H}{\partial x}\right)_{i\pm\frac{1}{2}}$ appearing in the left hand side of (5.37) yields:

$$\frac{1}{\Delta x}\left\{K_{i+\frac{1}{2}} \cdot \left[\frac{H_{i+1} - H_i}{\Delta x}\right] - K_{i-\frac{1}{2}} \cdot \left[\frac{H_i - H_{i-1}}{\Delta x}\right]\right\} - \left\{\frac{\partial}{\partial x}\left[K\left(\frac{\partial H}{\partial x}\right)\right]\right\}_i =$$

$$\left\{\frac{\Delta x^2}{24} \cdot \frac{1}{\Delta x}\left[K_{i+\frac{1}{2}} \cdot \left(\frac{\partial^3 H}{\partial x^3}\right)_{i+\frac{1}{2}} - K_{i-\frac{1}{2}} \cdot \left(\frac{\partial^2 H}{\partial x^3}\right)_{i-\frac{1}{2}}\right] + O(\Delta x^3)\right\} + \tag{5.38b}$$

$$\frac{\Delta x^2}{24}\left\{\frac{\partial^3}{\partial x^3}\left(K \frac{\partial H}{\partial x}\right)\right\}_i + O(\Delta x^4)$$

Using again equation (5.37) with $Y = K \frac{\partial^3 H}{\partial x^3}$ yields a Taylor development for the finite difference appearing on the right hand side above:

$$\frac{1}{\Delta x}\left[K_{i+\frac{1}{2}} \cdot \left(\frac{\partial^3 H}{\partial x^3}\right)_{i+\frac{1}{2}} - K_{i+\frac{1}{2}} \cdot \left(\frac{\partial^3 H}{\partial x^3}\right)_{i-\frac{1}{2}}\right] = \left\{\frac{\partial}{\partial x}\left(K \frac{\partial^3 H}{\partial x^3}\right)\right\}_i +$$

$$\frac{\Delta x^2}{24}\left\{\frac{\partial^3}{\partial x^3}\left(K \frac{\partial^3 H}{\partial x^3}\right)\right\}_i + O(\Delta x^4).$$

Plugging this identity into the right hand side of (5.38b) yields, on the right hand side of that equation:

$$\frac{\Delta x^2}{24} \cdot \left\{\frac{\partial}{\partial x}\left(K \left(\frac{\partial^3 H}{\partial x^3}\right)_i + \frac{\partial^3}{\partial x^3}\left(K \frac{\partial H}{\partial x}\right)_i\right\} + O(\Delta x^3)\right.$$

Now observe that the left hand side of (5.38b) is just the one-dimensional truncation error function $T_1(H) = L_1(H)$.

Furthermore, the exact mid-nodal conductivities can be expressed as $K_{i+½} = \hat{K}_{i+½} - \delta K_{i+½}$, where $\delta K_{i+½}$ is the error due to the approximate evaluation of $K_{i+½}$ by $\hat{K}i+½ = \sqrt{K_i \cdot K_{i+1}}$. Thus, we obtain an intermediate formulation of the truncation error $T(H) = \hat{L}(H) - L(H)$, shown below for just the first of three terms ($T = T_1+T_2+T_3$):

$$T_1(H) =$$

$$\frac{1}{\Delta x}\left\{\delta K_{i+½}\cdot\left[\frac{H_{i+1} - H_i}{\Delta x}\right] - \delta K_{i-½}\cdot\left[\frac{H_i - H_{i-1}}{\Delta x}\right]\right\} + \frac{\Delta x^2}{24}\cdot\left\{\frac{\partial}{\partial x}\left[K\,\frac{\partial^3 H}{\partial x^3}\right]_i + \right.$$

$$\left. + \frac{\partial^3}{\partial x^3}\left[K\,\frac{\partial H}{\partial x}\right]_i\right\} + O(\Delta x^3).$$

In order to complete this evaluation, we need again to replace finite differences by differentials. Thus, using again (5.36) in the above equation gives:

$$T_1(H) =$$

$$\frac{1}{\Delta x}\left\{\delta K_{i+½}\cdot\left(\frac{\partial H}{\partial x}\right)_{i+½} - \delta K_{i-½}\cdot\left(\frac{\partial H}{\partial x}\right)_{i-½}\right\} + \qquad (5.39)$$

$$\frac{\Delta x^2}{24}\cdot\frac{1}{\Delta x}\left\{\delta K_{i+½}\cdot\left(\frac{\partial^3 H}{\partial x^3}\right)_{i+½} - \delta K_{i-½}\cdot\left(\frac{\partial^3 H}{\partial x^3}\right)_{i-½}\right\} +$$

$$\frac{\Delta x^2}{24}\cdot\left\{\frac{\partial}{\partial x}\left(K\,\frac{\partial^3 H}{\partial x^3}\right)_i + \frac{\partial^3}{\partial x^3}\left(K\,\frac{\partial H}{\partial x}\right)_i\right\} + O(\Delta x^3).$$

Finally, we must use again a Taylor expansion technique to evaluate the "mid-nodal conductivity error" $\delta K_{i\pm\frac{1}{2}}$ that appears in equation (5.39). We have from previous definitions:

$$\begin{cases} \delta K_{i+\frac{1}{2}} = \hat{K}_{i+\frac{1}{2}} - K_{i+\frac{1}{2}} \\ \qquad = \sqrt{K(x_i)\cdot K(x_{i+1})} - K(x_{i+\frac{1}{2}}). \end{cases}$$

However, it will be more convenient to use the log-conductivity process $f(\underline{x}) = \ell n\, K(\underline{x})$ in order to evaluate the δK term. The mid-nodal log-conductivity is:

$$\begin{cases} \hat{f}_{i+\frac{1}{2}} = \ell n\, \hat{K}_{i+\frac{1}{2}} = \frac{1}{2}(f_i + f_{i+1}) \\ f_{i+\frac{1}{2}} = \ell n\, K_{i+\frac{1}{2}} \end{cases}$$

which leads to a simple expression for the *mid-nodal conductivity error*:

$$\delta K_{i+\frac{1}{2}} = K_{i+\frac{1}{2}} \cdot \{\exp(\frac{f_{i+1}+f_i}{2} - f_{i+\frac{1}{2}}) - 1\}. \qquad (5.40)$$

Furthermore, we obtain by Taylor expansion of f_i and f_{i+1} around $f_{i+\frac{1}{2}}$:

$$\frac{f_{i+1}+f_i}{2} = f_{i+\frac{1}{2}} + \frac{1}{8}(\frac{\partial^2 f}{\partial x^2})_{i+\frac{1}{2}} \cdot \Delta x^2 + O(\Delta x^4)$$

so that the mid-nodal conductivity error (5.40) can be expressed as:

$$\delta K_{i+\frac{1}{2}} = K_{i+\frac{1}{2}} \cdot \left\{ \exp\left[\frac{\Delta x^2}{8} \cdot \left(\frac{\partial^2 f}{\partial x^2} \right)_{i+\frac{1}{2}} + O(\Delta x^4) \right] - 1. \right\} \qquad (5.41)$$

This expression may be linearized under conditions which will be discussed in more detail in the sequel (these conditions require both Δx and σ_f to be small):

$$\boxed{\delta K_{i+\frac{1}{2}} \simeq K_{i+\frac{1}{2}} \cdot \left\{ \frac{\Delta x^2}{8} \cdot \left(\frac{\partial^2 f}{\partial x^2} \right)_{i+\frac{1}{2}} + O(\Delta x^4) \right\}.} \qquad (5.42)$$

Plugging this into the expression (5.39) for the truncation error, we obtain:

$$T_1(H) = \frac{\Delta x^2}{8} \cdot \frac{1}{\Delta x} \left[\left(\frac{\partial^2 f}{\partial x^2} \right) \cdot \left(K \frac{\partial H}{\partial x} \right) \right]_{i-\frac{1}{2}}^{i+\frac{1}{2}} +$$

$$\frac{\Delta x^2}{24} \cdot \left\{ \frac{\partial}{\partial x} \left(K \frac{\partial^3 H}{\partial x^3} \right) + \frac{\partial^3}{\partial x^3} \left(K \frac{\partial H}{\partial x} \right) \right\}_i + O(\Delta x^3).$$

The first term is a finite difference which can be evaluated by a Taylor expansion similar to that of equation (5.38):

$$\frac{1}{\Delta x} \left[\frac{\partial^2 f}{\partial x^2} \cdot K \frac{\partial H}{\partial x} \right]_{i-\frac{1}{2}}^{i+\frac{1}{2}} = \frac{\partial}{\partial x} \left[\frac{\partial^2 f}{\partial x^2} K \frac{\partial H}{\partial x} \right]_i + O(\Delta x^2).$$

This gives finally in closed form the truncation error function $T_1(H)$ evaluated at the grid point $x = x_i$ for the one-dimensional flow operator. The result is the same for $T_m(H)$ with $m = 1,2,3$ corresponding to the spatial operators $L_1 = \frac{\partial}{\partial x_1}(K\frac{\partial}{\partial x_1})$, etc. The total truncation error in the 3D case obtains by summing:

$$T(H) = T_1(H) + T_2(H) + T_3(H) \tag{5.43}$$

where:

$$\boxed{T_1(H)=\frac{\Delta x_1^2}{24}\left\{3\,\frac{\partial}{\partial x_1}\left[\frac{\partial^2 f}{\partial x_1^2}\cdot K\frac{\partial H}{\partial x_1}\right]+\frac{\partial}{\partial x_1}\left[K\frac{\partial^3 H}{\partial x_1^3}\right]+\frac{\partial^3}{\partial x_1^3}\left[K\frac{\partial H}{\partial x_1}\right]\right\}+O(\Delta x_1^3))} \tag{5.44}$$

and similar expressions hold for T_2 and T_3.

We may now use this result in order to solve the partial differential equation governing the "equivalent" solution error $\delta\tilde{H}$. Using the Δx-expansion (5.34) and plugging (5.44) into equation (5.35), we find that:

(i) At order zero, $\delta\tilde{H}^{(0)}$ satisfies the flow equation with boundary conditions $\delta\tilde{H}^{(0)}=0$ so that $\delta\tilde{H}^{(0)}$ vanishes identically over the whole domain.

(ii) Since $T(H)=O(\Delta x^2)$, all the odd-order error terms $\delta\tilde{H}^{(1)}$, $\delta\tilde{H}^{(3)}$, etc., vanish.

(iii) The leading error term is $\delta\tilde{H}^{(2)}$, second order in Δx.

In the three-dimensional case, the leading order solution error can be expressed formally as:

$$\delta\tilde{H} = \sum_{i=1}^{3} \left\{ \delta\tilde{H}_m^{(2)} \cdot (\Delta x_m)^2 + O(\Delta x_m^4) \right\} \qquad (5.45)$$

and $\delta\tilde{H}$ is governed by equation (5.35) without the term $T(\delta\tilde{H}) = O(\Delta x^4)$, which can be neglected. This leads to a tractable governing equation for the solution error, of the form:

$$\frac{\partial}{\partial x_m}\left[K(\underline{x}) \cdot \frac{\partial(\delta\tilde{H})}{\partial x_m} \right] = - T(H) + O(\Delta x^4) \qquad (5.46)$$

Upon substituting the expression for T(H) given in (5.43)-(5.44), we obtain an equation which can be broken into three spatial component of the leading error term ($\delta\tilde{H}_m^{(2)}$, m = 1,2,3). For convenience, we show here only the term on the left-hand side of the equation corresponding to the first spatial direction m = 1:

$$\Delta x_1^2 \cdot \frac{\partial}{\partial x_1}\left[K(\underline{x})\frac{\partial E_1}{\partial x_1}\right] + \frac{\Delta x_1^2}{24}\left\{3\frac{\partial}{\partial x_1}\left[\frac{\partial^2 f}{\partial x_1^2} \cdot K\frac{\partial H}{\partial x_1}\right]\right.$$

$$\left. + \frac{\partial}{\partial x_1}\left[K\frac{\partial^3 H}{\partial x^3}\right] + \frac{\partial^3}{\partial x_1^3}\left[K\frac{\partial H}{\partial H_1}\right]\right\} + O(\Delta x_1^3) \tag{5.47}$$

where the notation $E_1 = \delta\tilde{H}_1^{(2)}$ was used. We also obtain similar expressions for E_2 and E_3. The total solution error is:

$$\delta\tilde{H} = E_1\Delta x_1^2 + E_2\Delta x_2^2 + E_3\,\Delta x_3^2 + O(\Delta x_m^4) \tag{5.48}$$

In the particular case where $\Delta x_1 = \Delta x_2 = \Delta x_3$ (equal mesh sizes in all directions), the full error equation becomes is simpler. Indeed we obtain the governing equation for $E = E_1 + E_2 + E_3$ by summing over spatial directions and observing that $\frac{\partial}{\partial x_m}(K\frac{\partial H}{\partial x_m})$ vanishes. Thus, we obtain finally the head solution error for a cubic mesh as follows:

$$\boxed{\delta\tilde{H} = E \cdot \Delta x^2 + O(\Delta x^4)} \tag{5.49}$$

where $E(x)$ is governed by the partial differential equation:

$$\boxed{\sum_m \frac{\partial}{\partial x_m}\left[K(x)\,\frac{\partial E}{\partial x_m}\right] = -\frac{1}{24}\left\{3\sum_m \frac{\partial}{\partial x_m}\left[\frac{\partial^2 f}{\partial x_m^2}K\frac{\partial H}{\partial x_m}\right] + \frac{\partial}{\partial x_m}\left[K\frac{\partial^3 H}{\partial x_m^3}\right]\right\} + O(\Delta x)} \tag{5.50}$$

with boundary conditions $E = 0$ on all boundaries.

Equation (5.50) gives explicitly the stochastic equation governing the head solution error. The first term on the right hand side is solely due to errors in the evaluation of mid-nodal conductivities, while the second term corresponds to errors in the evaluation of the flux divergence. Note that the right hand side depends solely on the exact solution H, which must be evaluated in order to find closed form stochastic solutions for the error $E(x)$. Our approach assumes that the remainder $O(\Delta x)$ can be neglected; this residual term could perhaps be loosened to $O(\Delta x^2)$ with more work.

5.2.2 Statistical analysis of the numerical head error

[a] <u>The spectrum of the head error:</u>

We now proceed to analyze the hydraulic head solution error $\delta\tilde{H} = E \cdot \Delta x^2$ in a stochastic framework. Following the assumptions of the spectral theory and the single realization approach, we postulate the equivalence of ensemble and spatial averages in the stochastic error equation (5.50). Furthermore, we use the first order spectral method to evaluate the statistical moments of the "exact" solution $H(x)$ appearing on the right hand side of (5.50). The same method will then be used to

obtain the first and second order moments of the head error itself. First, equation (5.50) is re-arranged in a more tractable form by using the relation:

$$K(\underline{x}) = K_G \, e^{f(\underline{x})}$$

where $f(\underline{x})$ is the perturbation of the ℓn K process around its mean ($\langle f \rangle = 0$). This leads to:

$$\sum_m \frac{\partial^2 E}{\partial x_m^2} + \frac{\partial f}{\delta x_m} \frac{\partial E}{\partial x_m} =$$

$$\sum_m \left\{ -\frac{3}{24}\left[\frac{\partial^3 f}{\partial x_m^3} \cdot \frac{\partial H}{\partial x_m} + \frac{\partial^2 f}{\partial x_m^2} \cdot \left[\frac{\partial H}{\partial x_m^2} + \frac{\partial f}{\partial x_m}\frac{\partial H}{\partial x_m}\right]\right] - \frac{1}{24}\left[\frac{\partial^4 H}{\partial x_m^4} + \frac{\partial f}{\partial x_m} \cdot \frac{\partial^3 H}{\partial x_m^3}\right]\right\} \tag{5.51}$$

Equation (5.51) should be interpreted as a stochastic equation in an infinite domain, i.e., with domain size much larger than correlation scales. The boundary conditions $E = 0$ suggest that the mean error should be approximately zero in order to be consistent with the stationarity-ergodicity assumptions. We will show that this is indeed the case for σ_f small. Our purpose will be to determine the standard deviation, or root-mean-square norm (RMS norm) of the solution error $\delta H = E \cdot \Delta x^2$, in an ensemble sense. The accuracy of the finite difference approximation will then be studied by analyzing the

behavior of $\sigma(\delta H)$ compared to $\sigma(H)$. The ratio of these two quantities indicates the relative amount of "numerical noise" compared to the physical noise in the solution.

In order to show explicitly how the forthcoming statistical analysis of error depends on σ_f being small, let us reproduce below the spectral perturbation solution of the exact flow equation (as in Chapter 3). We begin by expanding (5.53a) H(x) into a power series with σ_f as the "small parameter":

$$H(x) = H_0(x) + \sigma H_1(x) + \sigma^2 H_2(x) + \cdots$$

Similarly, we let:

$$f(x) = \sigma \cdot g(\underline{x})$$

or equivalently:

$$K(\underline{x}) = K_G \cdot e^{\sigma \cdot g(x)}$$

where $g(\underline{x})$ is a zero-mean Gaussian field with unit variance. The 3D flow equation (5.1) with stochastic conductivities can now be expressed as:

$$\frac{\partial^2 H}{\partial x_m \partial x_m} + \sigma \cdot \frac{\partial g}{\partial x_m} \cdot \frac{\partial H}{\partial x_m} = 0.$$

By plugging the expansion (5.52) into this flow equation, we obtain the infinite hierarchy of equations:

$$\left|\begin{array}{l} \dfrac{\partial^2 H_0}{\partial x_m \partial x_m} = 0 \\ \quad \vdots \\ \dfrac{\partial^2 H_1}{\partial x_m \partial x_m} + \dfrac{\partial g}{\partial x_m} \cdot \dfrac{\partial H_0'}{\partial x_m} = 0 \\ \quad \vdots \\ \text{(etc.)}. \end{array}\right. \qquad (5.53b)$$

The zero order term (H_0) could be interpreted as the linear mean head solution, satisfying $\dfrac{\partial H_0}{\partial x_m} = -J_m$, where J_m is the mean hydraulic gradient. Accordingly, the other terms in the expansion will have zero mean as can be easily checked. Assuming that the mean gradient is parallel to x_1, the first order term in the expansion satisfies:

$$\frac{\partial^2 H_1}{\partial x_m \partial x_m} = \frac{\partial g}{\partial x_1} \cdot J_1 .$$

Comparing this to the equation for head perturbations (h) as obtained in Chapter 3, it appears that H_0 and H_1 are related to the mean and first order perturbation of the head field as follows:

$$\begin{cases} \langle H \rangle = H_0(\underline{x}) = -\underline{J} \cdot \underline{x} \\ h = \sigma \cdot H_1 + O(\sigma^2) \end{cases} \qquad (5.54a)$$

Solving the equation for H_1 in Fourier space gives immediately the Fourier increment and spectral density of H_1:

$$dZ_{H_1} = i\, J_1 \frac{k_1\, dZ_g}{k^2} \quad (i = \sqrt{-1}))$$

$$S_{H_1} = J_1^2 \frac{k_1^2}{k^4} S_g \tag{5.45b}$$

We now use a similar expansion for the stochastic error in order to solve equation (5.51) by a first order spectral method:

$$E(\underline{x}) = E_0 + E_1 \cdot \sigma + E_2 \cdot \sigma^2 + \cdots \tag{5.55a}$$

By plugging the expansions for H and E into (5.51) with $f(\underline{x}) = \sigma\, g(\underline{x})$, we obtain a hierarchy of equations for E_0, E_1, etc., as shown below:

$$\left|\begin{array}{l} \sigma^0: \quad \sum_m \frac{\partial^2 E_0}{\partial x_m \partial x_m} = 0 \\ \\ \sigma^1: \quad \sum_m \left\{ \frac{\partial^2 E_1}{\partial x_m \partial x_m} + \frac{\partial g}{\partial x_m} \cdot \frac{\partial E_0}{\partial x_m} \right\} = \\ \\ \qquad -\frac{3}{24}\left[- J_1 \cdot \frac{\partial^3 g}{\partial x_1^3} \right] - \frac{1}{24} \sum_m \frac{\partial^4 H_1}{\partial x_m^4} \\ \sigma^2: \text{ (etc.).} \end{array}\right. \tag{5.55b}$$

Again, the zero order equation suggests that E_0 is the mean error, and that it must vanish identically over the infinite domain. This is consistent with the boundary conditions $E = 0$ for the finite domain case. It is also consistent with the fact that all higher order terms appear to have zero mean if $E_0 = \langle E \rangle$ holds. As a consequence, the equation governing the leading term (E_1) becomes:

$$\boxed{\frac{\partial^2 E_1}{\partial x_m \cdot \partial x_m} = \frac{1}{8} J_1 \frac{\partial^3 g}{\partial x_1^3} - \frac{1}{24} \frac{\partial^4 H_1}{\partial x_m^2 \cdot \partial x_m^2}} \qquad (5.56)$$

with implicit summation over repeated indices. Thus, the head error's leading order term in Δx and σ_f is given by:

$$\boxed{\delta H = E \cdot \Delta x^2 = E_1 \cdot \sigma_f \cdot \Delta x^2.} \qquad (5.57)$$

The two equations above show that the numerical error δH is proportional, as a first approximation, to the product $\sigma \Delta x^2$ times a stochastic term governed by a stochastic PDE independent of σ and Δx.

Furthermore, since equation (5.56) is linear, it can be solved easily in Fourier space, i.e., by using a spectral method as in Chapter 3. Plugging Fourier-Stieltjes representations for

E_1 and g in (5.56) yields:

$$k^2 \cdot dZ_{E_1} = \frac{3}{8} J_1 (ik_1)^3 dZ_g - \frac{1}{24} (k_m^2 \cdot k_m^2) dZ_{H_1}.$$

Using also equation (5.54) for dZ_{H_1}, this gives explicitly the Fourier component and spectrum of $E_1(x)$ in terms of the Fourier component and spectrum of $g(x)$:

$$dZ_{E_1}(\underline{k}) = -i \frac{J_1 k_1}{24} \left\{ 3 \frac{k_1^2}{k^2} + \frac{(k_m^2 \cdot k_m^2)}{k^4} \right\} dZ_g(\underline{k}) \tag{5.58}$$

$$S_{E_1}(\underline{k}) = \frac{J_1^2}{(24)^2} k_1^2 \cdot \left[3 \frac{k_1^2}{k^2} + \frac{(k_m^2 \cdot k_m^2)}{k^4} \right]^2 \cdot S_g(\underline{k})$$

where we used again implicit summation over repeated indices. Equivalently, the spectrum of δH can be obtained by multiplying S_{E_1} by Δx^4 and replacing S_g by S_f. Thus, equation (5.58) gives the spectral density of the "equivalent error" $\delta\tilde{H}(\underline{x})$.

Note that $\delta\tilde{H}(\underline{x})$ is a zero-mean random field defined in the continuous 3D space. However, we are only interested in the discrete error defined at the nodes of the finite difference grid. The restriction $\delta\hat{H}_1$ of the continuous process $\delta\tilde{H}(\underline{x})$ on the FD grid may be viewed as a lattice process, whose spectrum is identical to that of the continuous error within the range of wavenumbers:

$$0 \leq |k_m| \leq \pi/\Delta x_m \qquad (m = 1,2,3).$$

and zero outside this range. In other words, the fluctuations of $\tilde{\delta H}$ at scales smaller than the mesh size must be ignored. In particular, the variance of the discrete error process $\hat{\delta H}_i$ can be obtained by integrating the spectrum (5.58) up to wavenumbers $(\pi/\Delta x)$. This gives the final result we were looking for, namely the variance of the head error for the 3D flow problem:

$$\boxed{\text{var}(\hat{\delta H}_i) \simeq \left[J_i \cdot \frac{\Delta x^2}{24}\right]_o^2 \cdot \int_{0\leq k\leq \pi/\Delta x} k_i^2 \left[\frac{3k_i^2}{k^2} + \frac{(k_m^2 \cdot k_m^2)}{k^4}\right]^2 \cdot S_{ff}(\underline{k})\, d\underline{k}} \qquad (5.59)$$

[b] Head error in the one-dimensional case:

For ease of analysis, we focus first on the one-dimensional version of the flow equation. In this case, equation (5.59) becomes simpler:

$$\text{var}(\hat{\delta H}_i) \simeq \left[J_i \frac{\Delta x^2}{6}\right]^2 \cdot \int_{0\leq k\leq \pi/\Delta x} k_i^2\, S_{ff}(k_i)dk_i. \qquad (5.60)$$

It is interesting to note that, in one dimension, the variance of the head error is proportional to the spectral content of df/dx up to wavenumber $(\pi/\Delta x)$. For illustration, let us use the 1D

"hole-spectrum" log-conductivity as proposed by *Bakr et al. 1978*:

$$S_{ff}(k) = \sigma^2 \frac{\ell^3}{(\pi/2)} \frac{k^2}{(1+k^2\ell^2)^2}$$

$$R_{ff}(\xi) = \sigma^2 \, (1 - |\xi|/\ell) \cdot e^{-|\xi|/\ell}.$$

In this case, the spectral head solution is known to be stationary, with variance:

$$\mathrm{Var}(H) = \sigma_h^2 = J_1^2 \, \sigma_f^2 \, \ell^2.$$

On the other hand, observe that the first derivative of the f(x) process used in this example has a significant spectral content at large wavenumbers, up to the wavenumber cut-off ($\pi/\Delta x$) corresponding to the smallest scale of fluctuations sampled by the numerical grid. The integral in (5.60) can now be obtained by using the following identities from Gradshteyn and Rhyzik, 1980 (2.174 and 2.175):

$$\int \frac{u^4}{(1+u^2)^2} du = \frac{u^3}{1+u^2} - 3 \int \frac{u^2}{(1+u^2)^2} du$$

$$\int \frac{u^2}{(1+u^2)^2} du = -\frac{2u}{4(1+u^2)} + \frac{2}{4} \cdot \int \frac{du}{1+u^2}$$

$$\int \frac{du}{1+u^2} = \mathrm{arctg}\, u$$

whence the result:

$$\frac{\mathrm{Var}(\delta\hat{H}_i)}{\mathrm{Var}(H)} = \frac{2}{\pi}\left(\frac{1}{6}\right)^2 \cdot \left\{\left(\frac{\Delta x}{\ell}\right)^4 + \frac{\left(\frac{\Delta x}{\ell}\right)^3 + \frac{3}{2}\left(\frac{\Delta x}{\ell}\right)^5}{\left(\frac{\Delta x}{\ell}\right)^2 + 1} - \frac{3}{2}\left(\frac{\Delta x}{\ell}\right)^4 \cdot \mathrm{arctg}\left[\left(\frac{\Delta x}{\ell}\right)^{-1}\right]\right\}.$$

For $\Delta x/\ell \ll 1$, this gives a simple expression for the ratio of the standard deviations of the head numerical error and head solution in one dimension:

$$\frac{\sigma(\delta\hat{H}_i)}{\sigma(H)} \simeq \frac{1}{6}\sqrt{2/\pi}\left(\frac{\Delta x}{\ell}\right)^{3/2} + O\left(\frac{\Delta x}{\ell}\right). \qquad (5.61)$$

The most striking feature in this equation is that the "noise-to-signal ratio" appears independent of the input log-conductivity variance, and increases as a fractional power of the resolution ($\Delta x/\ell$), rather than the usual $O(\Delta x^2)$ behaviour for deterministic problems. The relative error is 2.5% for a resolution 1/3, and 7% for a resolution 2/3.

The simplicity of the 1D case allows us to study in some detail, the effect of the behaviour of f(x) at small scales or large wavenumbers. If a smoother process with 1D exponential covariance is used, equation (5.61) becomes:

$$\frac{\sigma(\delta\hat{H}_i)}{\sigma_f \cdot J \cdot \ell} \simeq \frac{1}{6} \cdot \left(\frac{\Delta x}{\ell}\right)^2$$

Alternatively, using the 1D Band-Pass Self-Similar Spectrum of Chapter 4, with $\Delta x \lesssim \ell << L$, we obtain:

$$\frac{\sigma(\delta\hat{H}_i)}{\sigma_H} \simeq \frac{1}{6} \left(\frac{\Delta x}{\sqrt{L\ell}}\right)^2 .$$

On the whole, these results indicate that the relative numerical error goes to zero with $(\frac{\Delta x}{\lambda})$ faster in the case of a smooth $\ell n\, K$ process, compared to the case where $\ell n\, K$ has significant variability at the smallest scales. This observation can be stated formally as follows:

$$\frac{\sigma(\delta\hat{H}_i)}{\sigma(H)} \simeq \frac{1}{6} \left(\frac{\Delta x}{\lambda}\right)^p, \qquad 0 < p \lesssim 2 \qquad (5.62)$$

where λ is a typical correlation scale, and p is equal to 3/2 for the "noisy" $\ell n\, K$ process with hole-exponential covariance, while $p = 2$ for smoother processes.

[c] Head error in the three-dimensional case:

We will now see that the relative numerical error $\sigma(\delta H)/\sigma(H)$ follows a similar behavior in the case of 3D flow.

The variance of the three-dimensional head error given in equation (5.59) will be computed below for two different isotropic log-conductivity spectra: a "noisy" spectrum, and a "smooth" spectrum. The isotropic 3D Markov spectrum is a good candidate for a "noisy" random field, since it is non-differentiable in the mean-square sense (the variance of df/dx_m is infinite!). This noisy random field was used for all the stochastic single-realization flow simulations to be described in Chapters 6 and 7. On the other hand, the 3D Hole-Gaussian spectrum (used by Vomvoris, 1986, for analysis of stochastic solute transport) appears as a good candidate for a "smooth" random field, being infinitely differentiable.

To obtain a closed form result for Var(δH) in equation (5.59) requires the evaluation of complicated three-dimensional Fourier integrals. The details of this evaluation are given in *Appendix* 5A for the case of the Markov spectrum (noisy field), and *Appendix* 5B for the case of the Hole-Gaussian spectrum (smooth field). These appendices also develop the spectral solution for the head field in order to obtain Var(H). The root-mean-square relative numerical error on the head field is given below, respectively, for the "noisy" case and the "smooth" case:

(a): Noisy input: $$\frac{\sigma(\delta H)}{\sigma(H)} = \frac{\sqrt{3}}{12} \left(\frac{\Delta x}{\lambda}\right)^{3/2} \left(1 + 0\left(\frac{\Delta x}{\lambda\pi}\right)\right)^{1/2} \tag{5.63}$$

(b): Smooth: $$\frac{\sigma(\delta H)}{\sigma(H)} \leq \frac{\sqrt{5}}{8} \left(\frac{\Delta x}{\ell}\right)^2$$

This result confirms the behavior observed earlier in the 1D case. Indeed, equation (5.63) shows that the relative head error is proportional to the grid resolution $\Delta x/\lambda$ with a power 3/2 far the "noisy" 3D Markov log-conductivity spectrum, and with a power 2 for the "smooth" 3D Hole-Gaussian spectrum. Note that the λ-scale stands for the integral correlation scale of the Markov spectrum, while the ℓ-scale is a typical fluctuation scale for the Hole-Gaussian spectrum (not the integral scale). Both log-conductivity fields are assumed isotropic, and the grid has equal mesh size in all three directions. A generalization of these results to the case of anisotropic inputs and rectangular grids would be of great interest.

Equation (5.63) can be used to compute the leading order term of the relative head error in specific cases (note that the inequality in 5.63b becomes equality as the resolution ratio goes to zero). It appears that the relative error on the head field is quite small and fairly independent on the type of log-conductivity field (noisy or smooth), at least for a reasonably fine grid. Indeed, the error is only 3% in both cases

for a resolution equal to 1/3. This indicates that the hydraulic head fluctuations can be accurately resolved by the finite difference flow simulator with moderate grid resolution. The situation is not so good when it comes to evaluating the Darcy flux vector, as will be seen shortly.

5.2.3 Numerical error on the flux vector

[a] Relation between flux error and head error:

The next step of the truncation error analysis focuses on the error in evaluating the flux vector q_m by the centered FD scheme (5.31). This error is defined as:

$$\delta\hat{q}_m = \hat{q}_m - q_m \qquad (m = 1,2,3). \qquad (5.64)$$

The FD solution $\hat{q}_m$ was expressed in terms of head differences in equation (5.31). Dropping the spatial direction index (m) for convenience, equation (5.31) gave an expression of the form:

$$\hat{q}(x_{i+\frac{1}{2}}) = \hat{K}_{i+\frac{1}{2}} \cdot \frac{\hat{H}_{i+1} - \hat{H}_i}{\Delta x}. \qquad (5.65)$$

Thus, the flux error evaluated at the mid-nodal points of the grid is simply:

$$\delta\hat{q}\ (x_{i+\frac{1}{2}}) = -\left\{K_{i+\frac{1}{2}} \cdot (\frac{\partial H}{\partial x})_{i+\frac{1}{2}} - \hat{K}_{i+\frac{1}{2}} \frac{\hat{H}_{i+1} - \hat{H}_i}{\Delta x}\right\} \quad (5.66)$$

Now, we can use our previous results in order to evaluate explicitly the flux error. Indeed, let:

$$\hat{K}_{i+\frac{1}{2}} = K_{i+\frac{1}{2}} + \delta\hat{K}_{i+\frac{1}{2}}$$

$$\hat{H}_i = H_i + \delta\hat{H}_i .$$

Plugging these expressions in equation (5.66) yields:

$$\delta\hat{q}(x_{i+\frac{1}{2}}) = -\left\{K_{i+\frac{1}{2}} (\frac{\delta H}{\delta x})_{i+\frac{1}{2}} - K_{i+\frac{1}{2}} \cdot \frac{H_{i+1} - H_i}{\Delta x}\right\}$$

$$-\left\{\delta\hat{K}_{i+\frac{1}{2}} \cdot \frac{H_{i+1} - H_i}{\Delta x}\right\} - \left\{k_{i+\frac{1}{2}} \frac{\delta\hat{H}_{i+1} - \delta\hat{H}_i}{\Delta x}\right\}. \quad (5.67)$$

Using a Taylor expansion, the first term in braces takes the form:

$$-\left\{-K_{i+\frac{1}{2}} \frac{\Delta x^2}{24} (\frac{\partial^3 H}{\partial x^3})_{i+\frac{1}{2}} + 0\ (\Delta x^4)\right\}.$$

To evaluate the second term, we use our previous finding concerning the mid-nodal conductivity error $\delta K_{i+\frac{1}{2}}$ in (5.42). This gives:

$$- \left\{ K_{i+\frac{1}{2}} \frac{\Delta x^2}{8} \left(\frac{\partial^2 f}{\partial x}\right)_{i+\frac{1}{2}} + 0\ (\Delta x^4) \right\}.$$

Finally, we evaluate the third term by using a Taylor expansion of $(\delta\hat{H}_{i+1} - \delta\hat{H}_i)$ with $\delta\hat{H}_i$ replaced by the equivalent $\delta\hat{H}(x)$ head error. This gives:

$$- \left\{ K_{i+\frac{1}{2}} \frac{\Delta x^4}{24} \left(\frac{\partial^3 (\delta\tilde{H})}{\partial x^3}\right)_{i+\frac{1}{2}} + 0\ (\Delta x^4) \right\}.$$

Reassembling these terms in (5.67) leads to the following expression for the equivalent flux error $\delta\tilde{q}(\underline{x})$ defined in continuous space:

$$\delta\tilde{q}(\underline{x}) =$$

$$K(\underline{x}) \cdot \left\{ \frac{\Delta x^2}{24} \left[\frac{\partial^3 H}{\partial x^3} + 3 \frac{\partial^2 f}{\partial x^2} \cdot \frac{\partial H}{\partial x} \right] + 0\ (\Delta x^4) \right\}$$

$$- K(\underline{x}) \cdot \left\{ \frac{\partial}{\partial x} \delta\tilde{H} + \frac{\Delta x^2}{24} \frac{\partial^3}{\partial x^3} \delta\tilde{H} + 0\ (\Delta x^4) \right\}.$$

The head error $\delta\tilde{H}$ is known from previous results. Recall that the leading order term in σ_f and Δx was:

$$\delta\tilde{H} = E \cdot \Delta x^2 = E_1\ \sigma_f \Delta x^2$$

where $E_1(x)$ is known from equation (5.58). Similarly, the exact solution $H(x)$ is known from previous spectral perturbation results. In particular, we have:

$$\frac{\partial H}{\partial x} = - J + \sigma_f \cdot H_1 + O(\sigma_f^2)$$

where the zero-mean $H_1(x)$ process is known from equation (5.54). Note also that $f(x) = \sigma_f\, g(x)$. Plugging these expressions into the equation for $\tilde{\delta q}_m(\underline{x})$, we obtain to the leading order in σ_f and Δx:

$$\tilde{\delta q}_m(x) = \sigma_f \cdot \Delta x^2 \cdot K_G \exp(\sigma \cdot g) \left\{ \frac{1}{24} \frac{\partial^3 H_1}{\partial x_m^3} + \frac{\partial E_1}{\partial x_m} \right\}. \qquad (5.68)$$

Finally, denoting h the perturbation $h = H - \langle H \rangle$, equation (5.68) can be expressed in a form which clearly shows the dependence of δq on the exact head perturbation h and the head error δH:

$$\boxed{\tilde{\delta q}_m(x) = K(\underline{x}) \cdot \left\{ \frac{\Delta x^2}{24} \frac{\partial^3 h}{\partial x_m^3} + \frac{\partial}{\partial x_m} \tilde{\delta H} \right\}} \qquad (5.69)$$

where the first term is due solely to errors in the evaluation of the mid-nodal conductivities. A similar expression holds for the error $\tilde{\delta G}_m$ in the hydraulic gradient $G_m = \frac{-\partial H}{\partial x_m}$:

$$\boxed{\tilde{\delta G}_m = - \frac{\partial}{\partial x_m} \tilde{\delta H}.} \qquad (5.70)$$

[b] Spectrum of head gradient and flux errors:

The random field errors (5.69) and (5.70) are perfectly determined from previous results (head spectrum and head error spectrum). For the hydraulic gradient error, we obtain the following statistics spectrum:

$$\left|\begin{array}{l} \langle \delta\tilde{G}_m \rangle = 0 \\ \\ dZ(\delta\tilde{G}_m) = \frac{\Delta x^2}{24} J_i \cdot k_m k_i \left[\frac{3k_i^2}{k^2} + \frac{(k_n^2 \cdot k_n^2)}{k^4}\right] \cdot dZ_f \\ \\ S(\delta\tilde{G}_m) = \left(\frac{\Delta x^2}{24}\right) \cdot J_i^2 \cdot (k_m k_i)^2 \left[\frac{3k_i^2}{k^2} + \frac{(k_n^2 \cdot k_n^2)}{k^4}\right]^2 \cdot S_{ff} \end{array}\right. \qquad (5.71)$$

with implicit summation over repeated indices. In order to determine the statistics of the flux error $\delta\tilde{q}$ in a similar fashion, we need to linearize the conductivity in the following fashion:

$$K = K_G\, e^{\sigma \cdot g} \simeq K_G\, (1 + \sigma g + \cdots).$$

Now, we recognize that this approximation may be poor if σ_f is larger than unity. Also, recall that a similar approximation was made for the mid-nodal conductivity error δK in (5.41) and (5.42). For the time being, we postpone discussing the possible

inaccuracy of this linearization approximation.

Using the latter approximation, we obtain a tractable expression for the stochastic flux error to the leading order in σ_f and Δx:

$$\tilde{\delta q}_m(x) = K_G\, \sigma_f\, \frac{\Delta x^2}{24} \left\{ \frac{\delta^3 H_1}{\partial x_m^3} + 24\, \frac{\partial E_1}{\partial x_m} \right\}. \qquad (5.72)$$

Finally, since the spectra of H_1 and E_1 are known from previous results, this gives the required statistics of the flux error vector in closed form:

$$\begin{cases} \langle \tilde{\delta q}_m \rangle = 0 \\ dZ_{\tilde{\delta qm}} = K_G \cdot J_1 \cdot \frac{\Delta x^2}{24} \left\{ \frac{k_1 k_m^3}{k^2} + k_1 k_m \left[\frac{3k_1^2}{k^2} + \frac{(k_n^2 \cdot k_n^2)}{k^4} \right] \right\} \cdot dZ_f \\ S_{\tilde{\delta qm}} = (\text{term above squared}) \cdot S_{ff}. \end{cases} \qquad (5.73)$$

The first equation indicates that the mean flux error is null in an infinite domain; the second equation gives the complex Fourier-Stieltjes increment of the stochastic flux error vector, and the third equation indicates how the diagonal components of its spectral density tensor can be obtained. Recall that S_{ff} is the known spectrum of the input log-conductivity field.

[c] Head gradient and flux errors in the 1D case:

In order to facilitate the analysis, we now focus on applications of previous results in the *1D case*. The spectral densities of the random field errors $\delta\tilde{G}$ and $\delta\tilde{q}$ shown in (5.71) and (5.73) become much simpler in one dimension. Dropping the *tilde* (~) sign for convenience, we obtain in this simple case:

$$S_{\delta G}(k) = \left(\sigma_f J_1 \frac{4}{24} \Delta x^2\right)^2 \cdot k^4 \frac{S_{ff}(k)}{\sigma_f^2} \qquad (5.74)$$

$$S_{\delta q}(k) = \left(\sigma_f J_1 K_G \frac{5}{24} \Delta x^2\right)^2 \cdot k^4 \frac{S_{ff}(k)}{\sigma_f^2} \qquad (5.75)$$

Now, the variances $\mathrm{Var}(\delta G)$, $\mathrm{Var}(\delta q)$, can be obtained by integrating the above spectra up to wavenumber $\pi/\Delta x$ as explained previously. It turns out that these variances are proportional to the variance of the *second* derivative of the log-conductivity field (after elimination of fluctuation scales smaller than the mesh size). For illustration, the relative numerical errors for the 1D Hole-Exponential Covariance log-conductivity ("noisy" field) were obtained in closed form. The result is shown below:

$$\frac{\sigma(\delta G)}{(\sigma_f J)} = \frac{\sigma(\delta q)}{(\frac{5}{4}\,\sigma_f J K_G)} \simeq$$

$$\sqrt{2/3} \cdot \pi/6 \cdot (\frac{\Delta x}{\ell})^{1/2} \cdot (1+0(\frac{\Delta x}{\pi \ell}))^{1/2} \qquad (5.76)$$

The most remarkable feature in equation (5.76) is that the order of accuracy on both the *head gradient* and the *flux* vector, drops by one power of $(\frac{\Delta x}{\ell})$ compared to the order of accuracy on the head (equation 5.61). This is exactly the same behaviour as in the case of deterministic, spatially smooth conductivities. However here, the order of accuracy is *fractional* rather than integer, and less than unity.

[d] Head gradient and flux errors in the 3D case:

We now proceed to develop similar expressions for the relative error on the numerical flux and head gradient solutions in the case of three-dimensional flow with statistically isotropic conductivity fields. We focus specifically on the "noisy" Markov spectrum of log-conductivity, which is the input spectrum actually used in the numerical simulations of groundwater flow presented in Chapter 6. The present analysis of the flux error, due to truncations in the finite difference scheme, is of particular interest for assessing the feasibility of accurate numerical simulations of three-dimensional flow and solute transport in heterogeneous media. Indeed, the spatial

fluctuations of the flux (or groundwater velocity) are responsible for the mechanical dispersion of convected solutes. It seems clear that an accurate simulation of the velocity field is a prerequisite for obtaining reliable simulations of the convection-dispersion mechanism in groundwater contamination problems.

In order to compute the *relative* error on the flux and head gradient vectors in the root-mean-square sense, one needs to compute both the variance of errors and the variance of the physical quantities themselves (using known solutions). The variance of numerical errors can be computed by integrating the error spectra given in equations (5.71) and (5.73), up to wavenumbers $|k_i| \leq \pi/\Delta x$ $(i = 1,2,3)$ in three-dimensional Fourier space. This can be expressed as follows, for the flux error vector $(Y_m = \delta q_m)$ as well as the head gradient error vector $(Y_m = \delta G_m)$:

$$\text{Var}\ (Y_m) = \iiint\limits_{\substack{0 \leq |k_i| \leq \pi/\Delta x \\ i=1,2,3}} S_{Y_m}(\underline{k}) \cdot d\underline{k} \qquad (5.77)$$

The result of integration is given in Appendix 5.C, using the flux error spectrum (5.73) and the head gradient error spectrum (5.71). Note that the domain of integration in (5.77) was approximated as $0 \leq k \leq \pi/\Delta x$, where k is the radial wavenumber.

In addition, the error variances obtained in Appendix 5.C take particularly simple forms for moderate-to-small grid resolution ratio:

$$R = \frac{\Delta x}{\pi\lambda} << 1 \ ,$$

which is precisely the case of most interest for applications (say $\Delta x/\lambda \lesssim 1/2$ in practice). Thus, we will generally assume $R << 1$ in what follows.

On the other hand, the variances of the flux and head gradient random fields are known from the first order spectral solutions developed in Chapter 3 (see equations 3.21-3.24). These results are reproduced below for convenience:

$$\begin{cases} \mathrm{Var}(q_1) = \{\sqrt{\frac{8}{15}}\, K_G \sigma_f J\}^2 \\ \mathrm{Var}(q_2) = \{\sqrt{\frac{1}{15}}\, K_G \sigma_f J\}^2 \\ \mathrm{Var}(q_3) = \{\sqrt{\frac{1}{15}}\, K_G \sigma_f J\}^2 \end{cases}$$

and

$$\begin{cases} \mathrm{Var}(G_1) = \{\sqrt{\frac{3}{15}}\, K_G \sigma_f J\}^2 \\ \mathrm{Var}(G_2) = \{\sqrt{\frac{1}{15}}\, K_G \sigma_f J\}^2 \\ \mathrm{Var}(G_3) = \{\sqrt{\frac{1}{15}}\, K_G \sigma_f J\}^2 \end{cases}$$

The final result shown below in terms of relative root-mean-square errors was obtained by computing the ratio $\mathrm{Var}(\delta Y)/\mathrm{Var}(Y)$ and taking the square root (see Appendix 5.C). The relative numerical error on the flux vector is:

$$\boxed{\begin{aligned} m = 1: \quad & \frac{\sigma(\delta q_1)}{\sigma(q_1)} \simeq \frac{2.42\pi}{24} \left(\frac{\Delta x}{\lambda}\right)^{1/2} \\ m = 2,3: \quad & \frac{\sigma(\delta q_m)}{\sigma(q_m)} \simeq \frac{1.82\pi}{24} \left(\frac{\Delta x}{\lambda}\right)^{1/2} \end{aligned}} \tag{5.78}$$

and the relative numerical error on the head gradient is:

$$\boxed{\begin{aligned} m = 1: \quad & \frac{\sigma(\delta G_1)}{\sigma(G_1)} \simeq \frac{3.51\pi}{24} \left(\frac{\Delta x}{\lambda}\right)^{1/2} \\ m = 2,3: \quad & \frac{\sigma(\delta G_m)}{\sigma(G_m)} \simeq \frac{1.39\pi}{24} \left(\frac{\Delta x}{\lambda}\right)^{1/2} \end{aligned}} \tag{5.79}$$

where λ is the integral correlation scale of the 3D isotropic Markov log-conductivity field.

5.2.4. Summary and discussion

We have evaluated the root-mean-square norm of the

finite difference solution error on the random head and flux fields, , for both 1D and 3D saturated flow. The three-dimensional results of equations (5.78) and (5.79) confirm the behaviour already observed in the one-dimensional case (equation 5.76). In both cases, it appears that the order of accuracy on the flux vector is $O(\Delta x/\lambda)^{1/2}$ when a "noisy" log-conductivity field is used. Recall that the order of accuracy on the hydraulic head was $O(\Delta x/\lambda)^{3/2}$ in the "noisy" case, and $O(\Delta x/\lambda)^2$ in the case of a "smooth" log-conductivity field. In spite of the fact that we did not compute the order of accuracy for the flux in the case of a "smooth" 3-*dimensional* log-conductivity field, there is little doubt that the order of accuracy will be $O(\Delta x/\lambda)$ in that case, as shown previously for the 1-dimensional case. These findings are summarized in Table 5.1.

The most important conclusion to be drawn from the truncation error analysis developed above, is that the centered finite difference scheme is a *consistent approximation* of the stochastic flow equation, even when the log-conductivity is a "noisy" random field, such as the non-differentiable 3D Markov field. Here, "consistency" means convergence in the mean-square sense of the finite difference solution to the exact solution as

Table 5.1

Order of Accuracy of the Stochastic Finite Difference Approximation (the "Noisy" and "Smooth" Random Fields were Defined in the Text, with Examples for 1 and 3-Dimensional Flow).

ORDER OF ACCURACY $O(\Delta x/\lambda)^p$	NOISY ℓn K FIELD	SMOOTH ℓn K FIELD
Hydraulic Head (H)	p = 3/2	p = 2
Head Gradient (G_m)	p = 1/2	p = 1
Flux Vector (g_m)	p = 1/2	p = 1

Table 5.2

Relative Numerical Error on the Hydraulic Head and Flux Vector, in the Case of the 3D Isotropic Markov Log-Conductivity Spectrum ("Noisy" Random Field)

GRID RESOLUTION ($\Delta x/\lambda$)	1/10	1/3	1/2
Hydraulic Head (H)	0.5%	3%	5%
Longitudinal flux(q_1)	10. %	17%	22%
Transverse Flux (q_2, q_3)	8. %	14%	17%

the grid resolution ($\Delta x/\ell$) goes to zero. The length scale ℓ represents the integral correlation scale or some otherwise defined fluctuation scale, of the log-conductivity field.

Furthermore, it is worth noting that the present analysis leads to explicit estimates of the leading order term of the numerical error for the variables of interest, particularly the hydraulic head and the flux vector. Table 5.2 gives the relative numerical errors on the hydraulic head and on the different components of the flux vector in the case of the 3D Isotropic Markov spectrum of log-conductivity. Recall that the relative error was defined as the ratio of the standard deviation of the numerical error δY, versus the standard deviation of the variable of interest Y. For a moderate grid resolution, such as the value 1/3 used in the numerical experiments of Chapter 6, the error on the head appears to be fairly small (3%), while the error on the flux vector is significant but still acceptable (less than 20%).

Another important finding from the truncation error analysis is that the *relative* errors $\sigma(\delta Y)/\sigma(Y)$ appear to be independent of the log-conductivity standard deviation σ_f. Therefore, as σ_f decreases, the *absolute* precision on σ_H and σ_{q_i} will improve, but not the *relative* precision. The latter

can only be improved by using higher grid resolution. This finding will be useful for the design of numerical experiments aimed at obtaining accurate second order moments of the flow field for comparison with spectral solutions (Chapter 6).

In order to put these findings in proper perspective, it may be useful to recall the various assumptions that were used in our analysis of accuracy of the finite difference scheme. First of all, note that the solution errors were evaluated by using a statistical root-mean-square norm, or standard deviation of the random field error. It was assumed that the computational domain is large enough that these statistics can be viewed equivalently as ensemble or spatial averages, provided also that both the solution and the error be stationary and ergodic. Without these assumptions, no simple closed form results could be obtained.

Second, we emphasize the fact that the results of error analysis were obtained by using a double-expansion in terms of $\Delta x/\lambda$ and σ_f, respectively. Thus, the closed form results obtained above give only the *leading order term* of the root-mean-square error, with respect to the "small parameters" $\Delta x/\lambda$ and σ_f. In particular, note that a linearization of the type:

$$e^f \simeq 1 + f + \cdots$$

was needed in order to include the effect of inaccurate evaluation of mid-nodal conductivities by the geometric mean weighting scheme (see equations 5.40 and 5.41). This particular linearization could lead eventually to underestimating the error on the flux vector $\sigma(\delta q_i)$ for large values of σ_f. However, the same type of linearization was also used to evaluate $\sigma(q_i)$, so that the *relative* error $\sigma(\delta q_i)/\sigma(q_i)$ could be less dependent on the linearization approximations, even for large values of σ_f. As an indication, the leading order errors obtained in this section are thought to be fairly representative of the actual finite difference solution errors for values of σ_f up to 1-1.5 and $\Delta x/\lambda$ up to 0.5.

In addition, it is worth noting that the present error analysis did not include finite size effects and/or non-stationary behavior of the stochastic solution (as will necessarily occur to some degree for finite domain simulations). Neither did it include the sampling errors that will occur when computing single realization flow statistics by spatial averages (rather than ensemble averages). These effects will be discussed in Chapter 6.

Finally, it should be noted that several different kinds of errors actually occur in the numerical solution procedure: *truncation errors*, *matrix solution errors*, and *round-off errors*. The *truncation errors* were defined and analyzed in this section. *Solution errors* arise from the approximate solution of the finite difference system (e.g., by an iterative method). *Round-off errors* are due to the finite precision of digital computers, usually 32 digits on mainframes and 64 digits on recent supercomputers, and they accumulate more or less rapidly depending on the type of algorithm used (certain linear system solvers are more stable to round-off errors than others). It seems important to ascertain that solution errors and round-off errors will be minimal. Indeed, the fine grid resolution required for the case at hand leads to very large finite difference systems that may be difficult to solve accurately. This will be the subject of a forthcoming section, where we will focus on a particular type of preconditioned iterative solution method (the SIP solver).

By way of closing remarks, let us mention that the proposed approach of evaluating finite difference truncation errors for a stochastic partial differential equation appears to be new, in view of the available literature on numerical analysis. Our method, based on Taylor expansions, stochastic

linearization, and Fourier space representations, could be used to analyze mesh resolution requirements for more general stochastic partial differential equations, such as parabolic convection-diffusion equations. The present work could also be used to obtain a more complete picture of the discretization errors, such as the spatial correlation structure of the errors on the hydraulic head and on the flux vector. Other possible applications could involve the study of higher order finite difference methods, finite element methods, pseudo-spectral methods, and multigrid methods for stochastic PDE's with random coefficients.

5.3 Iterative Matrix Solver and Convergence Analysis for Linear Random Flow Problems

5.3.1 Review of iterative and preconditioned matrix solvers

Our survey of the literature (Table 5.3) clearly showed that large sparse matrix systems can be solved more efficiently with iterative solvers than with direct solvers such as Choleski factorization or Gauss substitution. One of the many examples of the superiority of iterative solvers, even for relatively modest

Table 5.3 List of Matrix Solvers and References

Solvers	Comments	Applications to Subsurface Hydrology and Related Problems
1 Point Jacobi	Explicit type; $O(n^2)$ iterations.	seldom used
2. Gauss Seidel	Weakly implicit; $O(n^2)$ iterations.	seldom used
3. SOR	Accelerated Gauss-Seidel O(n) iterations with optimal parameter, else $O(n^2)$ iterations.	Reisenauer et al., 1981 Bjordammen and Coats, 1969 Stone 1968
4. LSOR	Line - SOR, implicit along lines; requires optimal parameter for O(n) convergence.	Bjordammen and Coats, 1969 Freeze, 1971, Cooley, 1974
5 ADI	Implicit along each direction alternatively; requires near-optimal sequence of parameters.	Stone 1968 Weinstein et al., 1969 Bjordammen and Coats, 1969 Watts 1971; Cooley, 1974; Trescott and Larson 1977; Kershaw, 1978
6. ICCG	First order accurate Incomplete Choleski factorization, with Conjugate Gradient iterations (no parameter required).	Kershaw 1978 Gambolati, 1979 Kuiper 1981 and 1987 Gambolati and Perdon 1984
7. SIP	Second order accurate strongly implicit LU factorization (requiring sequence of parameters); Picard iterations.	Stone 1968, Weinstein et. al. 1969, Cooley 1974 and 1983, Trescott, 1975, Trescott and Larson 1977, McDonald and Harbaugh, 1984, Kuiper 1981 and 1987.
8. Gauss Elimination or Choleski Factorization	Fully implicit, direct solver (non-iterative).	Neuman and Davis 1983 Yeh and Luxmoore 1983

size problems, can be found in Gambolati and Perdon (1984). Our own experience with a Gauss elimination solver adapted to a banded Galerkin coefficient matrix confirmed this view. The CPU time for this direct solver was proportional to $N^{7/3}$ for steady state 3D groundwater flow on a cubic domain discretized into N elements. On a Vax 11/782 machine, the CPU time was about one hour for N = 4000 elements. However, solving a problem of size N = 64000 would have required *one month* of CPU time on the same machine, and larger problems on the order of 1 million elements could not be solved in reasonable amounts of time even on recent supercomputers such as Cray 2. The storage would be likewise prohibitive, being proportional to $N^{5/3}$ (number of equations N times the matrix bandwidth $N^{2/3}$). Thus, the storage requirement will be 10 Gigawords for large problems on the order of 1 million elements. In comparison, the central memory of the Cray 2 is currently about 250 Megawords.

The major disadvantage of using direct solvers for the solution of large linear systems lies in the fact that the triangular matrices arising in the process of decomposition are not sparse, even though the coefficient matrix itself may be sparse. For the 7-diagonal finite difference matrix depicted in Figure 5.2, a Choleski factorization $A = LL^T$ yields a triangular matrix L with mostly non-zero elements within the half band of width $n^2 = N^{2/3}$ (for a cubic grid of size $n^3 = N$). As

mentioned above, this yields on the order of $N^{5/3}$ non-zero elements to be computed, compared to just 4N non-zero elements in the lower half of the original matrix. This kind of observation has led numerical analysts to develop a number of iterative solution methods based on approximate *sparse* decompositions of the original system matrix, such that the computational work per iteration and the required storage are both proportional to N.

The sparse iterative methods can be roughly classified with respect to the approximate decomposition method used. Most of the "classical" iterative solvers are based on an approximate *splitting* of the matrix (Point Jacobi, Gauss-Seidel, successive overrelaxation (SOR), alternate directions implicit (ADI)) while the more recent "fast" iterative solvers are based on an approximate *factorization* of the matrix (strongly implicit procedure, and incomplete Choleski-conjugate gradients). The reader is referred to Jacobs (1981) for a survey of iterative solvers according to the classification proposed above, and Evans (1981) for a review of matrix-splitting preconditioners. A number of other reviews and experimentations with matrix iterative methods can be found in the collections of papers edited by Schultz (1981), Evans (1983), and Birkhoff and Schoenstadt (1984). In addition, Table 5.3 gives a list of references concerning the use of iterative solvers for subsurface flow and analogous problems; some of these studies include

numerical experiments and comparisons between different types of solvers.

Another important distinction to be made among iterative solvers is, precisely, the type of iteration used to converge to the solution of the original matrix system. Apart from the ICCG solver, all the other solvers mentioned above are based on Picard-like iterations, including in particular the SIP method. These solvers can be briefly described as follows. Consider the linear finite difference system:

$$Ah = b \tag{5.80}$$

and suppose that an approximation M of matrix A has been found (M must be easier to invert than A). The simple manipulation shown below leads quite naturally to a Picard iteration scheme where the new system matrix M is by construction easier to invert than the original matrix A:

$$(M + A - M)\, h = b$$

$$Mh = b + (M-A)h$$

$$Mh^{m+1} = b + (M-A)h^{m}.$$

Furthermore, by substracting Mh^m to both sides, one obtains a "modified Picard" scheme that is presumably more stable with respect to round-off errors:

$$M \cdot \delta h^{m+1} = b - Ah^m$$

where $\delta h^{m+1} = h^{m+1} - h^m$, and m is the iteration counter. Finally, the iterations can be overrelaxed or underrelaxed by multiplying the right-hand side residual by a "relaxation parameter" ω:

$$\boxed{M \cdot \delta h^{m+1} = \omega \cdot (b - Ah^m)} \qquad (5.81)$$

It is worth noting that (5.81) is a consistent iteration scheme with respect to the original system (5.80), in the sense that the exact solution $h = A^{-1}b$ is obtained as $m \to \infty$, provided however that the iterations converge. Unfortunately, convergence is not necessarily guaranteed in the general case.

We now proceed to review various kinds of preconditioners (matrix M). The classical iterative solvers such as Jacobi and various versions of successive overrelaxation (SOR) are based on an approximate decomposition obtained by splitting A into lower triangular, diagonal, and upper triangular

matrices:

$$A = L + D + U$$

The approximate matrix M can be expressed in the general form:

$$M = D^{-1}(D+\tau_L L)(D+\tau_u U) \qquad (5.82)$$

When the preconditioner (5.82) is plugged into (5.81), one obtains some well known iterative solvers of the "matrix-splitting" kind:

Point Jacobi	($\tau_L = \tau_u = 0$, and $\omega = 1$)
Point Gauss-Seidel	($\tau_L = 1$, $\tau_u = 0$, and $\omega = 1$)
Point SOR	($\tau_L = 1$, $\tau_u = 0$, and $1 \leq \lvert\omega\rvert \leq 2$)
Symmetric SOR	($\tau_L = \tau_u = \tau$, and $\omega = \tau(2-\tau)$)

Similarly, the ADI solver can be viewed as an iterative method based on matrix-splitting decomposition. As an example, the Peaceman-Rachford version of ADI for two-dimensional finite difference systems (Peaceman and Rachford, 1955), can be expressed as follows:

$$A = - L_X - L_Y$$

$$(\tau I - L_X)(h^{M+\frac{1}{2}} - h^m) = b - Ah^m$$

$$(\tau I - L_Y)(h^{m+1} - h^{m+\frac{1}{2}}) = b - Ah^{m+\frac{1}{2}}$$

where the L_X and L_Y matrices correspond to the partial differential operators in the X and Y-directions, respectively. Note that each step of ADI is similar to the basic iterative scheme (5.81). The ADI method can be extended to three-dimensional finite difference systems (Douglas, 1962). More general alternate directions operator splitting methods have also been devised for the solution of weighted residual and collocation systems (Celia and Pinder, 1985).

The convergence properties of some of the classical iterative solvers reviewed above have been thoroughly analyzed in the literature (Varga 1962, Young 1971, Golub and Van Loan 1983). For instance, it has been shown that Jacobi and Gauss-Seidel require on the order of n^2 iterations to reach a given precision, where n is the unidirectional size of the grid (Laplace problem, in any number of dimensions). The SOR methods require only O(n) iterations if the optimal iteration parameter can be computed accurately. Unfortunately, this requires estimating the spectral radius of the SOR iteration matrix. For

complex problems such as the stochastic flow equation, it seems unlikely that the "optimal" relaxation parameter could be estimated accurately without dramatically increasing the total computational work. Thus, the number of iterations is likely to be $O(n^2)$ - rather than the O(n) behaviour predicted for *optimal* SOR.

This feature is presumably shared by the iterative ADI methods. The theory for ADI convergences remains incomplete, but the work of Peaceman and Rachford (1955) and Wachspress and Habeter (1960) shows that the optimal ADI-relaxation parameter is not a constant. These authors proposed a cyclic sequence of parameters based on the eigenspectrum of the L_X and L_Y matrices defined above. The truly optimal sequence is not known, except for very special forms of the governing equation, e.g., the heat equation with spatially separable conductivity $K(x,y)=K_x(x)\cdot K_y(y)$. More details can be found in Varga (1962) and Ames (1977), among others.

The Jacobi, SOR, LSOR, and ADI solvers can also be compared in a slightly different way as follows. First of all, let us point out key feature shared by all the iterative methods reviewed above: the interactions among nodal variables are *partially decoupled* through the iterative solution process. At each iteration step, the solution is computed by taking into

account the interactions among a certain group of nodal values, while the remaining nodes are treated explicitly (e.g., by retaining the values obtained at the previous iteration). The computational work per iteration is roughly the same for all solvers (on the order of N), but the "degree of implicitness" differs. In the Jacobi method, the solution at each node is computed explicitly with respect to all the remaining nodes; in the Gauss-Seidel and SOR methods, about half of the nodes are treated implicitly on average; and in the ADI method, the nodal values are coupled alternatively in the X and Y-directions, while the other direction is treated explicitly. Various devices have been proposed in the literature in order to increase the degree of implicitness, or coupling, of iterative solvers while still retaining the advantages of a sparse and easily invertible matrix approximation. Indeed, the SOR matrix-splitting can be generalized into line-SOR or more generally block-SOR splittings, which may increase the coupling along lines or among neighboring nodes (Evans, 1984).

However, our literature review indicates that the most efficient, or "strongly implicit", iterative solvers are those based on an approximate *factorization* of the original matrix, such as the SIP and ICCG solvers mentioned earlier. This was taken into account in the classification given in Table 5.3, where the solvers were listed according to their "degree of

implicitness", increasing from top to bottom (Jacobi, Gauss Seidel, SOR, LSOR, ADI, ICCG, SIP, and Direct Solvers). According to this classification, direct solvers are fully implicit (requiring only "one iteration") but their computational cost will be prohibitive for large systems, as explained earlier. The advantage of SIP and ICCG is that they are based on *strongly implicit*, yet *sparse* factorizations of the coefficient matrix. Thus, these two methods presumably converge faster than the solvers based on matrix-splitting, while the computational work per iteration remains on the order of N.

The numerical experiments published in the literature confirmed this view. Most of the references given in Table 5.3 above involved comparisons between LSOR, ADI, SIP, and ICCG. It appeared that ADI had the slowest convergence rate in most cases, or diverged in difficult cases such a those involving anisotropic conductivities. The LSOR solvers were reasonably efficient, provided alternate line-sweeping along different directions, but the SIP solver was usually more efficient for "difficult" problems involving heterogeneities and mild nonlinearity. In addition, the performance of SIP was not overly sensitive to the choice of its iteration parameters, whereas this was sometimes a critical issue for LSOR and ADI.

On the other hand, ICCG also appeared to be a powerful solver. The numerical experiments by Kershaw (1978) demonstrate the superiority of ICCG over ADI and LSOR for a laser fusion problem (transient diffusion equation with a radiation term). Other numerical experiments for heterogeneous confined and unconfined groundwater flow (Kuiper 1981) indicate that SIP with underrelaxation could be as efficient as ICCG for the mildly nonlinear case of unconfined flow, although ICCG usually converged faster for linear problems. In a more recent study Kuiper (1987) concludes in favor of ICCG over SIP. However, it is possible that a change in the details of implementation, especially for the nonlinear problems, could affect his conclusions. Furthermore, the SIP solver involves an adjustable sequence of iteration parameters (similar to ADI), and could also be underrelaxed to avoid divergence in difficult cases. On the other hand, ICCG does not depend on any extraneous iteration parameter. It is conceivable that the flexibility of SIP could be an advantage, rather than a drawback, when dealing with near ill-conditioned systems.

At any rate, it may be preposterous to draw definite conclusions here, since the numerical experiments mentioned above were limited to rather modest-size flow problems, below 10,000 nodes. The largest among those was the saturated-unsaturated numerical simulation by Freeze (1971) with the LSOR solver on an

8,000-node grid, but his study did not include comparisons with other solvers. Larger simulations can be found in the literature on numerical analysis, however, these focus typically on the solution of the Laplace or Poisson equations (constant coefficients). One of these studies, by Jacobs (1983), develops an adaptation of the SIP factorization in conjunction with conjugate gradient iterations, for comparison with ICCG methods. The conclusions, based on two-dimensional test problems up to 40,000 nodes, were in favor of ICCG over SIP-CG. However, we do not know of any numerical experiments with the *standard* SIP solver for problems of comparable size or larger..

We now focus our review on the theory of SIP and ICCG, since our search in the literature indicates that these solution methods may have the best potential for large finite difference systems. The idea of using sparse approximate factorization for preconditioned iterative solvers arose in 1968, when the *SIAM Journal on Numerical Analysis* published in the same issue three papers on the approximate factorization and iterative solution of multi-dimensional finite difference systems. The first, by Stone, described the *strongly implicit procedure* (SIP) based on an approximate, non-symmetric LU factorization of the symmetric coefficient matrix A, with a Picard iteration scheme to converge to the solution of the A-matrix system. Although Stone's paper (1968) concerned only two-dimensional 5-point finite difference

systems, the SIP method was subsequently extended to three-dimensional 7-point finite difference systems (Weinstein et al, 1969). The two other papers in the 1968 *SIAM Journal*, by DuPont, Kendall, Rachford and the companion paper by Dupont, developed a symmetric LL^T approximate factorization with Picard iterations to converge to the solution. The LL^T factorization they used was similar to an incomplete Choleski factorization where the matrix L is forced to have the same sparsity pattern as A, and furthermore the row-sums of A are conserved (see Gustaffson 1978, Jackson and Robinson 1985). Thus, the iterative solvers of Stone (1968) and Dupont et al. (1968) differed essentially in the method used to obtain an approximate factorization of the finite difference matrix.

It was not until 1977 that the symmetric incomplete Choleski factorization was used as a preconditioner for conjugate gradient iterations (Mejerink and Van der Vorst, 1977). This combination, known as incomplete Choleski-conjugate gradients (ICCG), has become quite popular due to the fast convergence of the CG iterations in the case of well conditioned (preconditioned) symmetric positive-definite systems (see Kershaw, 1978, among others). It is interesting to note that the "pure" conjugate gradients method devised by Hestenes and Stiefel (1952) was viewed in the early days as an exact solver, since the method was known to converge to the exact solution in at most N

iterations (N being the number of equations). However, the CG iteration did not converge fast enough to be competitive as an approximate large sparse matrix solver (N large). Thus, it should be kept in mind that the incomplete factorization is an essential ingredient of the ICCG method, required to ensure fast convergence of the conjugate gradient iterations.

On the other hand, it is worth noting that the conjugate gradients method, and the ICCG solver, can only be used to solve symmetric positive-definite systems. As a consequence, the conjugate gradients method cannot be used to accelerate the convergence of the SIP solver, as the latter is based on a non-symmetric LU factorization. This would seem to advantage the ICCG method, since the conjugate gradients iterations presumably converge faster than Picard iterations for well conditioned (preconditioned) systems. On the other hand, the non-symmetric SIP factorization appears to be a more accurate approximation of the original system matrix than the Incomplete Choleski factorization (respectively second order and first order in Δx: see Stone 1968 and Gustaffson 1978). This seems to advantage SIP, with a better preconditioner than ICCG.

Unfortunately, there does not appear to be any solid theoretical basis on which to compare the two methods. A formal theory of SIP convergence is still lacking, due to the complex

form of the LU factorization involved; there has not been much progress in this area since the indications given by Stone (1968) for 2-dimensional problems with constant coefficients. The theory for ICCG is more developed, but limited to constant or mildly variable coefficients. For instance, Gustaffson (1978) showed that the condition number of the iteration matrix for the Picard-Incomplete Choleski solver of Dupont, et al. (1968) was $O(n)$, compared to $O(n^2)$ for the condition number of the original matrix (Laplace equation, with n the unidirectional size of the grid). On the other hand, the number of iterations required to reach a given precision grows like the square-root of this condition number, both for the Picard and the conjugate gradient iterations (Gustaffson 1978, and Golub and Van Loan 1983). This yields for the number of ICCG iterations a relation of the form:

$$m \sim n^{1/2}$$

which indicates that the ICCG method could converge quite fast for large 3D systems. Indeed, the total size of the grid is $N = n^3$ in three dimensions, which yields:

$$m \sim N^{1/6} \qquad (5.83)$$

indicating a very slow growth of the number of iterations with grid size. Note however that the theoretical analysis that led

to (5.83) was based on a number of assumptions, including the restriction to mildly variable or constant coefficients. It seems more reasonable to postulate that, in the worst case, the number of iterations could grow like the unidirectional size of the grid, for the ICCG as well as the SIP methods, i.e.:

$$m \sim n$$

This gives finally a worst case estimate of the number of iterations required for convergence of both ICCG and SIP for three-dimensional systems with highly variable coefficients:

$$\boxed{m \sim N^{1/3}} \qquad (5.84)$$

In comparison, note that the non-optimal SOR method will not converge faster than $N^{2/3}$ iterations, even for mildly variable coefficients.

We have developed a Fortran implementation of the SIP solver during the initial stages of this research. Some of the details of this implementation will be described in the next section, and a number of numerical experiments for large random flow problems will also be presented in a forthcoming section. Because the results obtained with SIP were eventually found to be quite satisfactory, it was felt that developing the ICCG solver

was not necessary. However, there is no claim that ICCG could not perform as well or perhaps better than SIP in terms of computational work (more on this later).

5.3.2 Formulation of the strongly implicit procedure (SIP solver)

We now proceed to describe the algebraic details of the strongly implicit procedure. Recall that SIP is based on the Picard iteration scheme (5.81):

$$M \cdot \delta h^{m+1} = \omega \cdot (b - A h^{m}) \qquad [(5.81)]$$

where $M = LU$ is an approximate non-symmetric factorization of the system matrix A. In what follows, we analyse in some detail the SIP factorization for the case of the 7-diagonal symmetric coefficient matrix A, corresponding to the 7-point centered finite difference scheme in three dimensions (see Figures 5.1 and 5.2 above). The 3D version of SIP was exposed briefly by Weinstein et al. (1969), based on the 2D version previously developed by Stone (1968). Details on coding can be found for instance in McDonald and Harbaugh (1984). However, our particular implementation is exposed below.

The SIP factorization aims at obtaining a close approximation of matrix A in the form of a product of a lower and

an upper triangular matrix (L and U) that have the same sparsity patterns as the lower and upper parts of A (Figure 5.5). The difficulty is to find the matrices L and U such that the product M = LU is indeed close to A in some sense. Let us define E the error matrix:

$$E = M - A = LU - A \tag{5.85}$$

Obviously, many choices of L, U, and E are possible for a given matrix A, since the system (5.85) is undetermined. On the other hand, it is easily seen by inspection that E cannot be the zero-matrix since the system A = LU is overdetermined. We conclude that A cannot be *exactly* factored in the form LU.

The particular factorization devised by Stone and co-workers was obtained by writing explicitly the undetermined system (5.85) - with L and U as shown in Figure (5.5) - in a recursive form. The undetermined coefficients in this recursion were then obtained by manipulating the equations in such a way that the product LU appears equivalent to a non-symmetric, second order finite difference approximation of the governing partial differential equation. In two dimensions, Stone (1968) showed that LU corresponds to a 7-point symmetric FD approximation

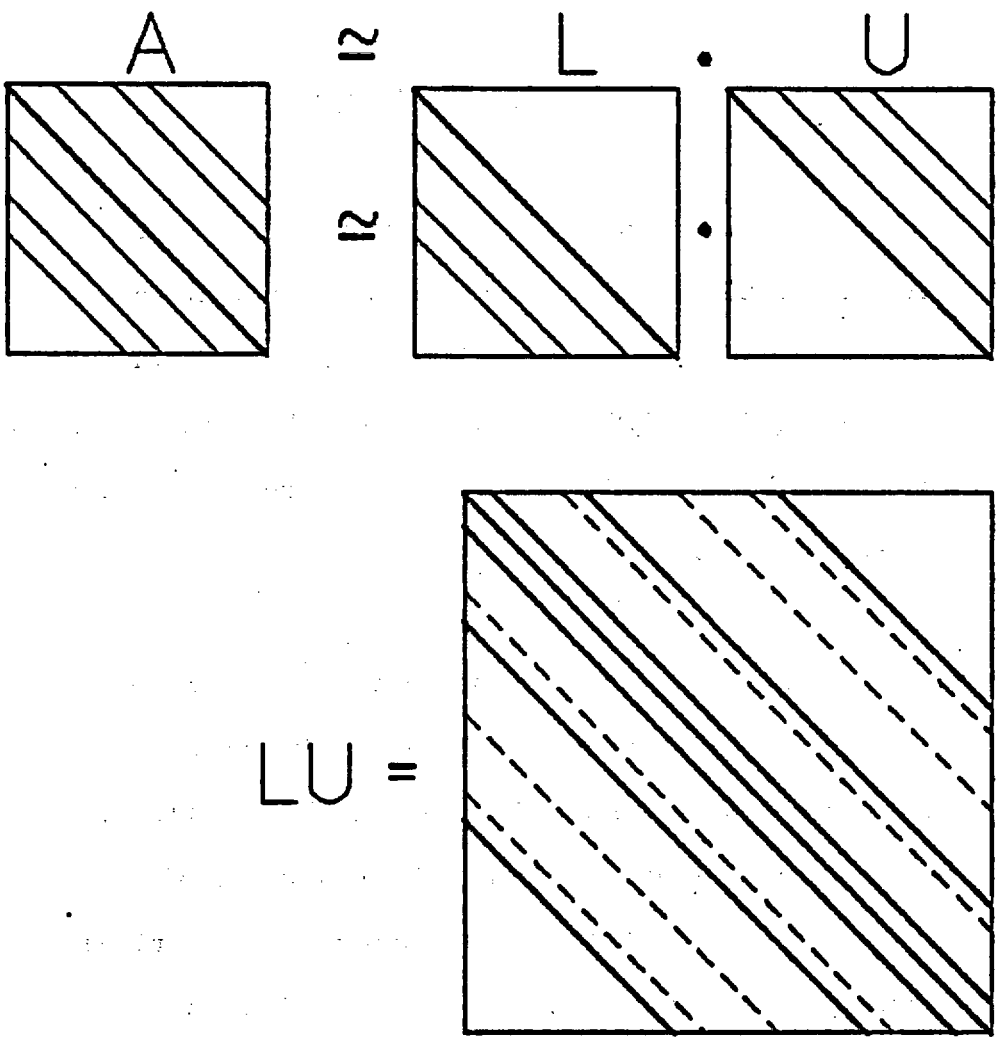


Figure 5.5 Schematic representation of the SIP approximate factorization (top), and structure of the product M = LU approximating A (bottom). The dashed lines indicate extra diagonals not present in the original matrix A.

(Figure 5.6, top) satisfying:

$$LU = A + O(\Delta x^2, \Delta y^2, \Delta x \Delta y).$$

In three dimensions, an analogous procedure devised by Weinstein et al. (1969) leads to a factorization LU that appears equivalent to a 13-point asymmetric FD approximation of the governing equation (Figure 5.6, bottom). Presumably, this corresponds to a second order approximation of the governing equation, as in the two-dimensional case.

However, it turns out that the LU factorization required an additional modification in order to ensure proper convergence of the iterations defined by (5.81). A new parameter τ was introduced (undetermined coefficient of the LU matrices) in such a way that the factorization described just above corresponds to the case $\tau = 1$. Note that τ is an "iteration parameter", distinct from the relaxation parameter ω appearing in equation (5.81). According to Stone (1968) and Weinstein et al. (1969), the best results were obtained when τ_m followed a cyclic sequence, with $0 < \tau_m < 1$. The proposed sequence of parameters is analogous to that of the ADI method. This sequence takes the form:

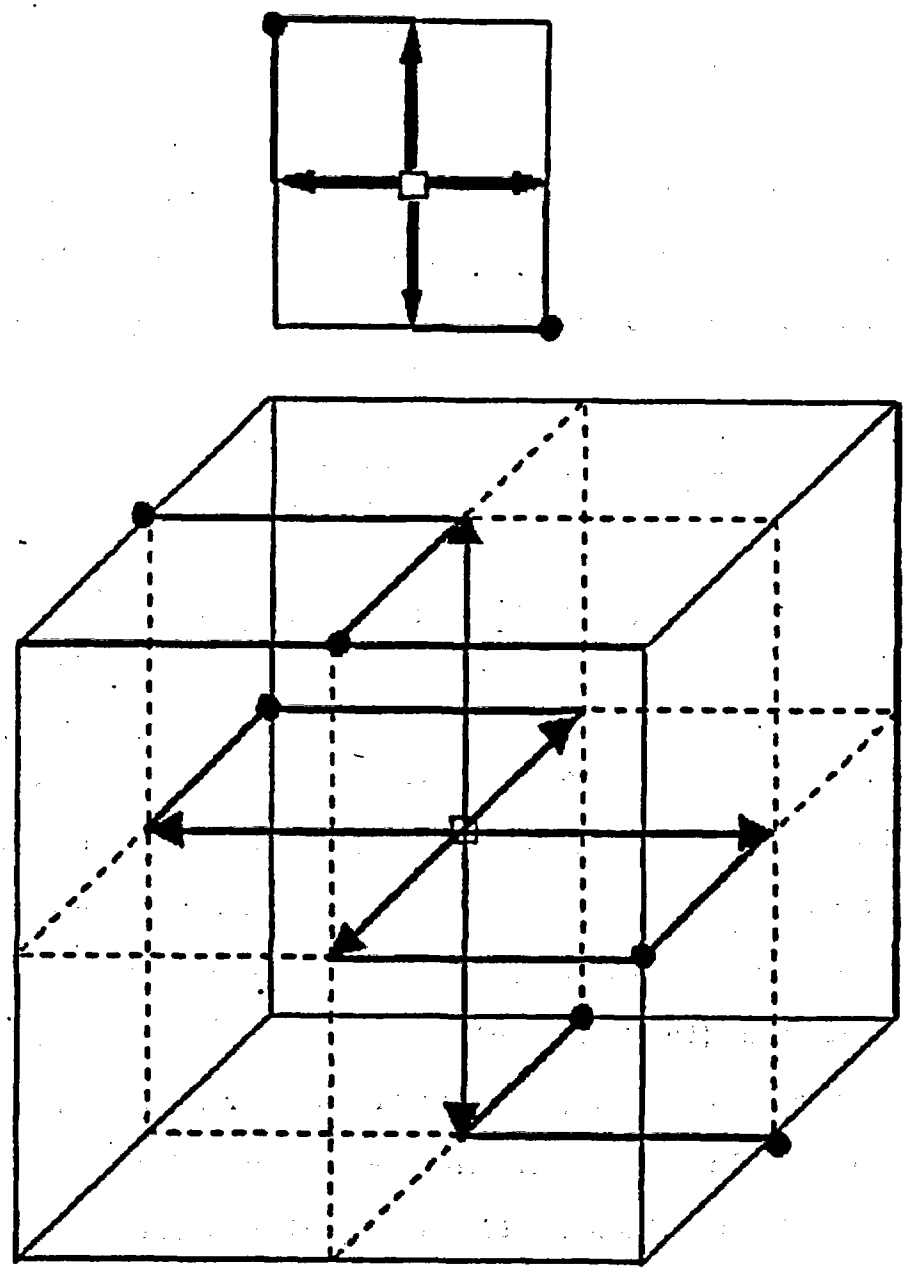


Figure 5.6 Asymmetric SIP molecule corresponding to the approximate LU factorization of the symmetric finite difference matrix A (top: 2D case; bottom: 3D case).

$$1 - \tau_m = (1-\tau_{max})^{\frac{m-1}{m_o-1}}, \quad m = 1,2,\cdots,m_o \qquad (5.86)$$

where m_o is the cycle length (taken equal to 4 in this work), and τ_{max} is the maximum value of τ over one cycle. Correcting the erroneous formula given by Stone (1968), τ_{max} is given by:

$$\tau_{max} = \underset{i,j,k}{\text{Min}} \left\{ \frac{\pi^2}{n_1^2 \rho_1}; \frac{\pi^2}{n_2^2 \rho_2}; \frac{\pi^2}{n_3^2 \rho_3} \right\}$$

$$\rho_1 = 1 + \left[\frac{K_{i,j+\frac{1}{2},k}}{\Delta x_2^2} + \frac{K_{i,j,k+\frac{1}{2}}}{\Delta x_3^2} \right] / \frac{K_{i+\frac{1}{2},j,k}}{\Delta x_1^2}, \text{ (etc.)}$$

where n_i represents the number of nodes along direction x_i. Note that τ_{max} appears to be spatially variable in the case of non-constant conductivities, so that the above formula requires some further modification. In the case of 3D flow with locally isotropic (but variable) conductivities, we propose a simple elimination of the conductivities appearing in ρ_1, etc., on the grounds that neighboring mid-nodal values should not differ much on a *high resolution grid*. This yields finally:

$$1 - \tau_{max} = \underset{i=1,2,3}{\text{Min}} \left\{ \frac{\pi^2 (\Sigma \Delta x_j^2 - \Delta x_i^2)}{2 n_i^2 \cdot \Sigma \Delta x_j^2} \right\} \qquad (5.87)$$

where $\Sigma\Delta x_j^2 = \Delta x_1^2 + \Delta x_2^2 + \Delta x_3^2$. Equations (5.86) and (5.87) completely define the cyclic sequence of SIP iteration parameters used in this work, both for the linear and nonlinear flow systems. The cycle length was $m_o = 4$ for all simulations. To avoid confusion, we stress again the fact that τ is distinct from the relaxation parameter ω appearing in the basic iteration (5.81). The parameter ω was introduced to provide additional flexibility to the SIP solver, and will turn out later to be useful.

It may be instructive to briefly explain the rationale behind the choice of a cyclic sequence of the τ-parameter. The geometric progression (5.86) was chosen intentionally in order to cover a wide range of values $0 \leq \tau_m \leq \tau_{max} < 1$, in such a way that the values taken by τ_m are mostly clustered near unity (note that τ_{max} is very close to one for large grids). Indeed, it turns out that for $\tau \simeq 1$, the iterative smoothing of the error is only effective within a narrow wavenumber band, so that several different values or τ near unity are needed in order to cover a wide enough band. In addition, the precise choice of τ_{max} is important for the success of the SIP method. According to equation (5.87), this parameter is close to one, but always less than one. Indeed, Fourier analysis shows that taking τ_{max} exactly equal to unity would lead to divergence of the SIP iterations.

In the case of the two-dimensional Laplace equation, Stone (1968) showed that, for any constant value of τ_m in the vicinity of unity, there will be an amplification of the Fourier modes of the error within some wavenumber range. On the other hand, Stone also showed that for any given Fourier mode, there exists a value of τ between zero and one that results in the decay of the amplitude of that mode. Values of τ near unity tend to decay the low wavenumber components of the error (although some other modes might be amplified!), while values of τ near zero tend to decay the high wavenumber modes quite rapidly. The sequence of τ_m defined above is therefore an important ingredient of the SIP method, without which very slow convergence or even divergence could occur. Unfortunately, there exists no proof that this sequence will guarantee optimal convergence of the SIP method in the general case. The ADI method encounters the same type of problem, as explained earlier.

Another important feature of the SIP factorization that might be altered to improve convergence is its directionality, arising from the asymmetry of the "SIP molecule" depicted in Figure 5.6 above. Indeed, the orientation of the SIP molecule could be changed by reordering the nodes in a different fashion. In three dimensions, we have enumerated eight possible ways to do so without changing the sparsity pattern of the system matrix. The standard node ordering scheme, which sweeps first through

$(+x_1)$, then $(+x_2)$, then $(+x_3)$, was used as the basic indexing scheme in our numerical code. However, the SIP algorithm can be manipulated to accommodate other node orderings, such as sweeping reversely through $(-x_1)$, then $(-x_2)$, then $(-x_3)$ in the negative directions. It has been claimed that implementing the SIP factorization with alternate ordering $(+x_1,+x_2,+x_3)/(-x_1,-x_2,-x_3)$ from one iteration to the next, resulted in improved convergence. However we have not observed any favorable effect of the alternate ordering strategy in our preliminary numerical tests. For completeness, note that there are two ways to implement the *alternate* SIP method; the one we have tried without success cycled the τ-parameter every other iteration, so that the same value of τ was used for alternate sweeps. According to Weinstein et. al. (1969), the other strategy which uses different values of τ over alternate sweeps may work as well or better. In any case, the numerical experiments shown in this work were all performed with the standard node ordering implementation of SIP (no alternate sweeps).

The algebra of the SIP factorization is tedious and will not be reproduced in detail here. The factorization must be recomputed at each iteration step because the factored matrices L and U depend on τ_m, which varies cyclically as explained before. Our particular implementation used a vector representation of each non-zero diagnonal line of A, L and U. We

obtained a nonlinear recursion for the coefficients of the L and U matrix, analogous to the Thomas algorithm for tridiagonal matrices, but much more complex (13 equations define the recursion). Unfortunately, this type of recursion does not appear to be fully vectorizable, except at the cost of increased storage requirements.

We have coded several implementations of the SIP solver in order to accommodate different computing environments. One of the variants is slower, but more flexible and requires less storage. Typically, the solution of a linear problem such as saturated flow with random field conductivities will require a storage of about 10 N words, where N is the total number of nodes. The number of equivalent additions performed at each iteration step is on the order of 50-100 per node, of which only a fraction is vectorizable. Thus, solving a single realization of the stochastic flow problem on a 3D grid on the order of 1 million nodes would require the availability of about 10 Megawords of central memory, and consume on the order of 100 MFLOP per iteration (1MFLOP = 1 million floating point operations).

In comparison, a recent supercomputer like the Cray 2 could run at 10 MFLOP/second (or several times faster) in scalar mode, and 100 MFLOP/second (or several times faster) in full

vector mode. The physical memory of the Cray 2 is currently 256 Megawords (with 64-bit words), but other supercomputers have only about 1-10 Mwords of central storage (Cray 1, Cyber 205). Nevertheless, storage should not be a real problem in the near future, since the current trend is towards larger physical memories on the order of the Gigaword (10^9 words).

We conclude that the numerical solution of random flow problems on grids on the order of 1 million nodes will be feasible at reasonable cost on current supercomputers having sufficient, direct access memory, *provided that the number of iterations required to reach an accurate solution be on the order of 1000 or less.* In this case, a solution for each single-realization problem could be reached in, say, no more than a few hours of CPU time. Thus, the key question is whether the SIP solver converges at a reasonable rate in the case of large, highly variable flow problems. We will examine this question below for the case of steady state flow in saturated media. The transient and nonlinear problem of unsaturated flow is of a different nature, and its numerical analysis is postponed to a later section.

5.3.3 Convergence analysis for large 3D random systems of saturated flow

In this subsection, we summarize the results of a number of numerical experiments conducted with the SIP solver, for large single-realizations of the stochastic groundwater flow equation in three dimensions (up to 1 million nodes). We begin by developing an approximate theory relating the unknown solution error to other "observables" such as the residual iteration error. This will serve as a basis to interpret the numerical experiments. Note that we focus here strictly on the *convergence* of the solver, and not on the physical meaning of the solutions themselves (see Chapter 6).

[a] Theoretical Analysis of Convergence:

We proceed to show that the convergence rate of the SIP solver is related to the spectral radius of the so-called iteration matrix, or "Jacobi matrix" J. Furthermore, we will also show that the observable "residual error" ($\hat{\epsilon}$) usually underestimates the true error (ϵ) by an amount which depends on the convergence rate. Let us start with the basic Picard iteration scheme (5.81):

$$LU(h^{m+1}-h^{m}) = \omega \cdot (b-Ah^{m}) \qquad [(5.81)]$$

and recall that LU is an *approximate* factorization of matrix A, which depends on the cyclic τ-parameter of equation (5.86). For simplicity here, we will ignore the fact that LU varies cyclically with respect to the iteration counter m. This should not be of consequence for the forthcoming analysis, e.g., the convergence rate can be interpreted as an average over the cycle length.

Let us now define a "residual error" which can be computed a posteriori by numerical experimentation:

$$\hat{\epsilon}^m = h^{m+1} - h^m \tag{5.88}$$

However, the "true error":

$$\epsilon^m = h^m - h \tag{5.89}$$

remains unknown, since the exact solution h is not known. Nevertheless, a recursive relation on ϵ^m is easily obtained by manipulating equation (5.81); this gives:

$$\epsilon^m = (J_\omega)^m \cdot \epsilon^o \tag{5.90}$$

where ϵ^o is the initial error (depending on the initial guess h^o), and J_ω is the iteration matrix:

$$J_\omega = I - \omega\,(LU)^{-1}\,A \tag{5.91}$$

Furthermore, it is easily seen, by manipulating again equation (5.81), that the residual error follows a geometric progression similar — but not identical — to equation (5.90):

$$\hat{\epsilon}^m = (J_\omega)^m \cdot \omega\,(h-(LU)^{-1}Ah^o) \tag{5.92}$$

These relations show that the SIP iterations will not converge unless the "norm" of the iteration matrix J_ω is less than one, in some sense to be precised later. In the case of an *exact* LU factorization, and taking $\omega = 1$, the iteration matrix becomes zero and the *exact* solution will be obtained after just one "iteration". In the case of the approximate SIP factorization, we expect that the rate of convergence be directly related to the accuracy of the LU factorization, which can be represented by the norm of J_ω in equation (5.91). This intuitive observation will be given a more precise meaning shortly.

Note also that equations (5.90) to (5.92) can be combined to relate the unknown error vector ϵ to the residual error vector $\hat{\epsilon}$ as follows:

$$\epsilon^m = (I - J_\omega)^{-1} \cdot \hat{\epsilon}^m \qquad (5.93)$$

This equation indicates that the true error may be much larger than the observed residual error when the norm of the iteration matrix J_ω is close to unity. Observe that J_ω depends on the chosen value of the relaxation parameter. The role of ω can be made more explicit by plugging equation (5.91) into (5.93). This yields:

$$\epsilon^m = \frac{1}{\omega} \cdot A^{-1} LU \cdot \hat{\epsilon}^m \qquad (5.94)$$

The role of the relaxation parameter now appears more clearly. Taking $\omega < 1$ (underrelaxation) does not seem a good strategy at first sight, since this will increase the ratio $\epsilon^m/\hat{\epsilon}^m$ according to equation (5.94). On the other hand, underrelaxation might be necessary in order to avoid divergence in the case where the LU matrix is not an accurate approximation of A (the norm of J_ω must be less than unity in equation 5.91). This suggests that there exists some optimal value of the relaxation parameter (ω_{opt}) which will maximize the convergence rate of the *true* error. It is conceivable that ω_{opt} be greater than one in certain "easy" cases, but more likely ω_{opt} will be less than unity for "difficult" problems characterized by a large condition number of the coefficient matrix.

Unfortunately, the relations obtained so far depend on the unknown inverse matrix A^{-1} involved in J_ω, so they cannot be used for an a priori analysis of convergence of the SIP solver in terms of the *true* error ϵ. However, it is possible to relate the norm or spectral radius of J_ω to the ϵ-convergence rate, at least approximately. In addition, comparing equation (5.92) to (5.90) shows that the ϵ- and $\hat{\epsilon}$-convergence rates should be the same. These remarks eventually lead to an explicit expression for the true error (ϵ) in terms of two "observable" quantities: the residual error ($\hat{\epsilon}$), and the $\hat{\epsilon}$-convergence rate. This is developed in more detail below.

Our starting point consists in obtaining a tractable expression for the true solution error (ϵ) in equation (5.93). Intuitively, this equation suggests an inequality of the form:

$$\|\epsilon^m\| \leq \frac{\|\hat{\epsilon}^m\|}{1-\|J_\omega\|}$$

A relation of this type was used for instance by Hageman and Young (1981) to design a stopping criterion for conjugate gradient iterations. In order to show that this is indeed a reasonable approximation of (5.93), we need to define some vector and matrix norms. The reader is referred to Householder (1964) for basic definitions and inequalities on matrix norms. The

facts directly relevant to the present work can be summarized as follows. A matrix norm must satisfy the usual properties of norms, including the triangular inequalities:

$$\begin{aligned} \|A + B\| &\leq \|A\| + \|B\| \\ \|AB\| &\leq \|A\| \cdot \|B\| \end{aligned} \tag{5.95}$$

One particularly useful matrix norm $\|A\|$ is defined in relation with the usual Euclidean norm $\|x\|$ for vectors, as follows:

$$\|A\| = \underset{x}{\text{Max}} \frac{\|Ax\|}{\|x\|} \tag{5.96}$$

Incidentally, it can be shown that $\|A\|$ is the square-root of the maximum eigenvalue of AA^T. Now, by using the second triangular inequality in (5.95) we obtain another useful inequality:

$$\|A^m\| \leq \|A\|^m \tag{5.97}$$

On the other hand, the spectral radius $\rho(A)$ of a matrix A is defined as its maximum absolute eigenvalue:

$$\rho(A) = \underset{i}{\text{Max}} \lambda_i(A) \tag{5.98}$$

This quantity turns out to be always smaller or equal to the Euclidean matrix norm:

$$\rho(A) \leq \|A\| \tag{5.99}$$

Finally, combining (5.97) and (5.99) yields another useful inequality:

$$\rho(A^m) \leq \|A^m\| \leq \|A\|^m \tag{5.100}$$

Let us now assume that the iterations do converge, and that the norm $\|J_\omega\|$ of the iteration matrix is less than unity (this also implies that its spectral radius $\rho(J_\omega)$ is less than unity). Equation (5.93) can then be approximated as follows. First, take the Euclidean vector norm on both sides of (5.93) to obtain:

$$\|\epsilon^m\| = \|(I-J_\omega)^{-1}\| \cdot \|\hat{\epsilon}^m\|.$$

Second, write a formal Taylor development of the matrix-valued function:

$$(I-J_\omega)^{-1} = I + J_\omega + J_\omega^2 + \cdots$$

and use the previous inequalities to obtain:

$$\|(I-J_\omega\|^{-1}) \leq 1 + \|J_\omega\| + \|J_\omega\|^2+\cdots$$

The series on the right-hand side converges since it was assumed that $\|J_\omega\| < 1$. Thus, we obtain finally:

$$\|(I-J_\omega^{-1})\| \leq \frac{1}{1-\|J_\omega\|}$$

which gives immediately the announced result:

$$\boxed{\|\epsilon^m\| \leq \frac{\|\hat{\epsilon}^m\|}{1-\|J_\omega\|}} \qquad (5.101)$$

On the other hand, applying previous matrix norm inequalities to equations (5.90) and (5.92) gives two more inequalities:

$$\begin{aligned} \|\epsilon^m\| &\leq \|J_\omega\|^m \cdot \|\epsilon^o\| \\ \|\hat{\epsilon}^m\| &\leq \|J_\omega\|^m \cdot \omega \, \|h-(LU)^{-1}Ah^o\| \end{aligned} \qquad (5.102)$$

Let us now define the convergence rate:

$$r = - \frac{d\ell n\ \epsilon}{dm} \tag{5.103}$$

and observe that the convergence rate of the residual error is identical to that of the true error if equality holds in both relations of (5.102). In that case, the convergence rate is given by:

$$r \simeq \ell n\ (1/\|J_\omega\|). \tag{5.104}$$

However, this can only be a rough approximation (hopefully on the safe side) since in fact equations (5.102) are *inequalities*. Other authors have used a slightly different argument that leads to a similar result with $\|J_\omega\|$ replaced by $\rho(J_\omega)$ in equation (5.104). This is reported for instance in Remson et al. (1971), following the work of Forsythe and Wasow (1960) and others. Indeed, expressing equation (5.90) in the basis (e_i) of independent eigenvectors of J_ω gives:

$$\epsilon^0 = \Sigma\ \alpha_i\ e_i$$
$$\epsilon^m = \Sigma\ \alpha_i\ \lambda_i^m\ e_i$$

and, in the case where the largest eigenvalue $\lambda_1 = \rho(J_\omega)$ dominates the others, we obtain approximately:

$$\epsilon^m \simeq [\rho(J_\omega)]^m \cdot \epsilon^o$$

This finally leads to replacing $\|J_\omega\|$ by $\rho(J_\omega)$ in equations (5.101) through (5.104). It should be mentioned that, although we know in general that $\rho(J) \leq \|J\|$, the cases where these two quantities are not close to each other are somewhat pathological. Here, we can only hope that the SIP iteration matrix J_ω is not pathological, i.e. that its spectral radius is approximately equal to its Euclidean matrix norm.

To conclude, the results of equations (5.101) - (5.104) finally lead to an upper bound on the "true" solution error, in the form:

$$\boxed{\|\epsilon^m\| \leq \frac{\|\hat{\epsilon}^m\|}{1-e^{-r}}} \qquad (5.105)$$

where ϵ is the true error vector, $\hat{\epsilon}$ is the residual error vector, and r is the convergence rate, which can be computed from:

$$\boxed{r \simeq -\frac{d\ \ell n\ \|\hat{\epsilon}^m\|}{dm}} \qquad (5.106)$$

Equations (5.105)-(5.106) are applicable in the case where the computed convergence rate, r, is a positive constant. In other words, the iterations must converge, and the number of iterations must be large enough that the convergence rate reaches an asymptotic value. Figure 5.7 shows three different situations that might arise in practice:

(i) The residual error norm $\|\hat{\epsilon}^m\|$ increases monotonically (on average over the cycle length) after a certain number of iterations: this is a sure sign that the method diverges.

(ii) The residual error norm $\|\hat{\epsilon}^m\|$ decreases monotonically but does not seem to reach a constant convergence rate: the method may converge with more iterations, or may not (possibly due to accumulation of round-off errors).

(iii) The residual error norm $\|\hat{\epsilon}^m\|$ decreases monotonically and reaches a constant convergence rate (straight line on a semi-log plot): the method clearly converges, and the final error $\|\epsilon^m\|$ can be evaluated a posteriori by using equation (5.105).

The methodology developed above proved to be useful in practice, especially for the solution of very large "random" flow

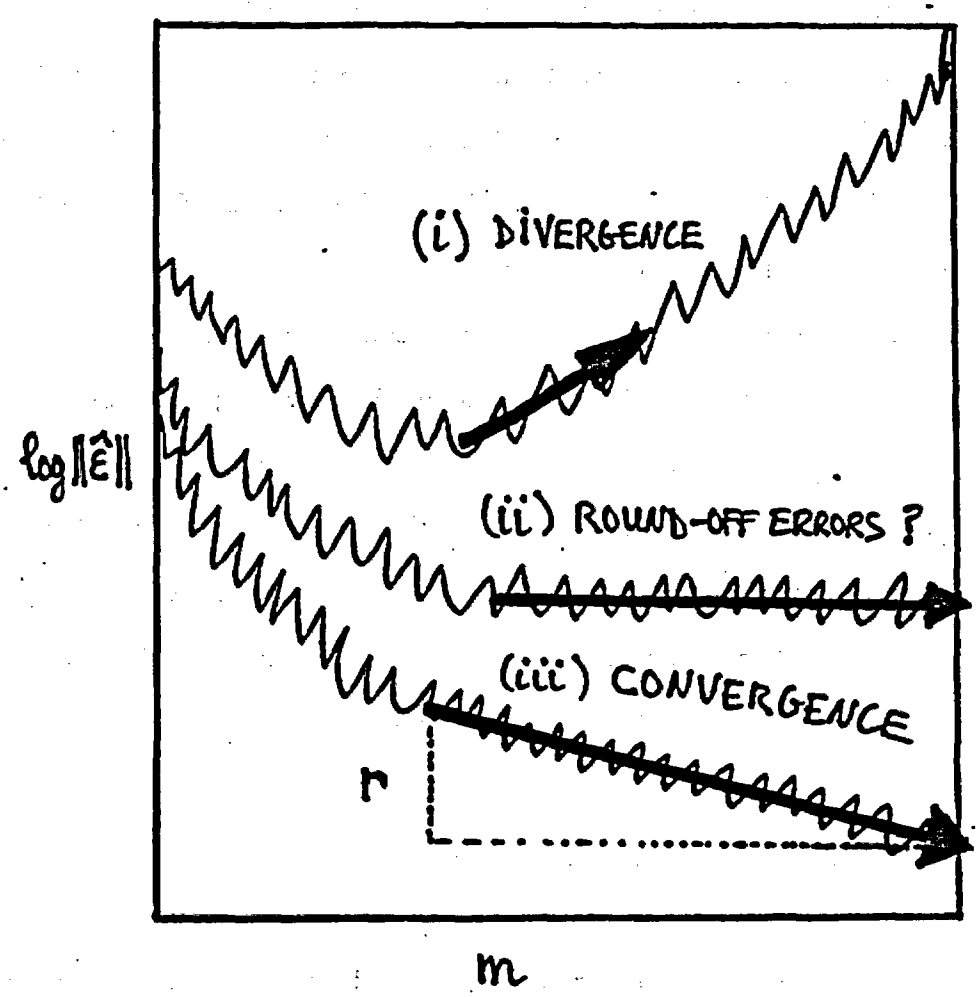


Figure 5.7 A posteriori analysis of convergence of the SIP solver: residual error norm versus number of iterations on a semi-log plot. In the convergent case (iii), the final residual error $\|\hat{e}\|$ and the convergence rate r can be used to estimate the true error $\|e\|$ as explained in the text.

problems. It was found that *underrelaxation* was needed in order to achieve convergence for such problems. As a consequence, the iterations converged quite slowly. Now, the simple relation in (5.105) shows that the *residual* error will largely underestimate the *true* solution error in these cases of slow convergence. Incidentally, let us mention that most empirical studies of iterative solvers in the literature use the residual error ($\hat{\epsilon}^m = h^{m+1}-h^m$), or else the right-hand side residual ($b-Ah^m$), in their evaluation of the solver's performance. The above analysis indicates that such information could be misleading, except perhaps for "academic" cases where the convergence rate is high enough that the residual error approximates well the true solution error. In the sequel, we analyze sequences of residual errors obtained from actual simulations, using equation (5.105) for a realistic evaluation of the performance of the SIP solver in terms of the "true" solution error.

[c] Numerical Experiments:

The SIP solver was applied to a model problem of saturated flow for a variety of grid sizes and different cases of log-conductivity variability. Briefly, the model problem was designed to simulate steady state saturated flow driven by a known global hydraulic gradient. The three-dimensional domain

was a rectangular prism, elongated in certain cases, cubic in other cases. The global hydraulic gradient was imposed along one of the axes of the rectangular prism (axis x_1) by using fixed head boundary conditions on two opposite faces (x_1=0 and x_1=L_1), while zero-flux conditions were imposed on all other lateral faces. With this configuration, x_1 represents the direction of the mean flow (longitudinal), while x_2 and x_3 are the axes transverse to the mean flow.

In the trivial case of a constant conductivity (σ_f=0), the exact solution in terms of the hydraulic head is a linear function of the longitudinal coordinate:

$$H^o(\underline{x}) = H(0) + \frac{H(L_1)-H(0)}{L_1} \cdot x_1 \qquad (5.107)$$

In the case of a random log-conductivity ($\sigma_f \neq 0$), equation (5.107) was used as the initial guess for the SIP solver. The log-conductivity was generated at each node of the finite difference grid by using the 3-dimensional Turning Band algorithm developed by Thomson, Ababou and Gelhar (1987). This method generates single-realizations of random fields, as explained earlier (Chapter 2). The particular random field used for the numerical experiments of this section is the 3D isotropic Markov field, whose spectrum was given in Table 3.1. Let us mention that other simulations were conducted with anisotropic Markov

log-conductivity fields. The numerical solutions obtained for isotropic as well as anisotropic fields will be analyzed more thoroughly at a later stage (Chapter 6).

Table 5.4 summarizes the results obtained with SIP for 4 test problems and several values of σ_f. The first row gives the size and geometry of the grid. Note that the largest problem (1 million nodes) was solved on a Cray 2 machine with 64-bit words, while the other problems were solved on a minicomputer, the Microvax 2, with 32-bit words. The mesh size was the same along all three directions, and equal to one third of the conductivity correlation scale for the largest problem (one half for the others).

The second row in Table 5.4 gives the standard deviation of the log-conductivity field (σ_f), which ranged between 1 and 2.3; these are fairly representative values in view of available field data (see Chapter 2). In addition, one of the simulations listed in Table 5.4 was for $\sigma_f = 0$, i.e. for constant conductivity (Laplace equation). The initial guess for the Laplace equation was taken to be a constant hydraulic head:

$$H(\underline{x}) = \frac{H(0)+H(L_1)}{2}$$

rather than the linear function (5.107) which is known to be the

Table 5.4

Convergence Rate of the SIP Solver:
Summary of Numerical Experiments for 3D Steady State
Saturated Flow with Statistically Isotropic Random Conductivities

Problem Label	A		B	C	D		
N	1 million		130,000	110,000	40,000		
n_i	(101; 101;101)		(51; 51; 51)	(250; 21; 21)	(80;21;22)		
σ_f	1.0	2.3	2.3	2.3	0.	1.	2.3
ω	2.5	.25	.50	.25	1.00	.50	.50
r	.0219	.0070	.0256	.0041	.0341	.0512	.0259
$(\Delta m_i)^*$	105.	330.	90.	570.	67.	45.	64.
$(s)^*$	1.2	0.4	0.7	0.3	0.9	1.4	0.7

(*) The "iteration increment" (Δm_i) and the "scaled convergence rate" (s) are defined in the text; note that Δm_i is the number of iterations required to decrease the true error by 1 order of magnitude.

exact solution of the Laplace equation. This particular test problem was included in order to examine the effect of conductivity variability over a wide range of σ_f values. In theory, the asymptotic convergence rate should not depend on the initial guess, but only on the size of the grid, the input conductivities, and certain iteration parameters.

The third row of inputs in Table 5.4 gives the relaxation parameter ω used for simulations (see equation 5.81). The method used to search for an optimal value of ω was rather elementary and empirical. The value $\omega = 1$ was tried first; if divergence or very slow convergence occurred, the simulation was started again with underrelaxation. Thus, the sequence of values of the relaxation parameter used successively in the search process was $\omega = 1$, $\omega = 0.50$, $\omega = 0.25$, and $\omega = 0.1$.

The last three rows of Table 5.4 display the asymptotic convergence rate (r) and two other quantities related to it. Recall that r is the rate of decrease of the logarithm of the residual error normal as defined by equation (5.106), which can be rewritten as:

$$r \simeq - \frac{d \ \ell n \ \|h^{m+1} - h^m\|}{dm} \qquad [(5.106)]$$

As explained earlier, this should be identical to the rate of convergence of the true error, $\|h-h^m\|$. The "iteration increment" Δm_1 was defined as the average number of iterations required to decrease the error by 1 order of magnitude:

$$\Delta m_1 = \frac{\ell n\ 10}{r} \tag{5.108}$$

Finally, the "scaled convergence rate" (s) was obtained by dividing the convergence rate by a factor *approximately* equal to the square root of the *condition number* of the coefficient matrix A. In fact, this condition number is known only for the special case $\sigma_f = 0$ (Laplace matrix), namely:

$$C_0 = \frac{4}{3} \frac{(n_1)^2+(n_2)^2+(n_3)^2}{\pi^2} \tag{5.109}$$

where n_i is the unidirectional size of the grid (number of nodes) in the direction x_i.

The condition number given by (5.109) was obtained as follows. First of all, the condition number of a symmetric matrix is defined as the ratio of the maximum versus minimum absolute eigenvalue of the matrix. The eigenspectrum of the Laplace matrix is easily obtained by solving the Laplace eigenproblem:

$$(\nabla^2-\lambda)\ v = 0$$

for $v(\underline{x})$, and by plugging the corresponding eigenvectors $v(i_1\Delta x_1,\ i_2\Delta x_2,\ i_3\Delta x_3)$ into the matrix eigenvalue problem:

$$(A-\lambda I)v = 0.$$

In the case where $\Delta x_1 = \Delta x_2 = \Delta x_3$, this gives the eigenvalue spectrum:

$$\lambda(\underline{k}) = \frac{K_G}{\Delta x^2} \cdot \sum_{i=1,2,3} \left[2 \sin \frac{k_i \pi}{2(n_i+1)}\right]^2$$

where $k_i = 1,2,\cdots n_i$. The final result (5.109) was obtained by computing the ratio of the extreme eigenvalues ($C_o = \lambda_{max}/\lambda_{min}$) assuming $n_i >> 1$, and the quantity "s" on the last row of Table 5.4 was defined as:

$$s = C_o^{-1/2} \cdot r \qquad (5.110)$$

Thus, a constant value of s across the last row of Table 5.4 would indicate that the convergence rate is proportional to $C_o^{1/2}$, which is itself approximately proportional to the unidirectional size of the grid along the direction of maximum elongation (largest number of nodes). To avoid confusion, recall that C_o is the condition number of the Laplace matrix ($\sigma_f = 0$), not of the

random matrix ($\sigma_f \neq 0$).

The major results that emerge from Table 5.4 are the following. First of all, it appears that underrelaxation was usually required in order to achieve convergence. Second, convergence was fairly slow in general, since the number of iterations required to decrease the error by 1 order of magnitude (Δm_i) was on the order of one hundred, up to several hundred iterations in the most difficult cases. Furthermore, it appears that the required number of iterations was roughly proportional to the unidirectional size of the grid as defined earlier. This can be seen by comparing problems A,B,C for $\sigma_f = 2.3$. On the other hand, the required number of iterations usually increased with σ_f for a given grid size (see problems A and D for $\sigma_f = 1$ and 2.3).

However, the sequence of convergence rates obtained for problem D with $\sigma_f = 0$, 1 and 2.3, indicates that the influence of σ_f on convergence might be fairly complex. If one accepts the conjecture that the convergence rate of SIP is proportional to $C_A^{1/2}$, where C_A is the condition number of the *random* conductivity matrix A, then the results of problem D suggest that C_A decreases with σ_f at low values of σ_f, and increases eventually for larger values. We have observed this kind of behavior for a very small matrix with random

conductivities distributed independently of each other, using a direct perturbation analysis to obtain a "geometric mean" condition number. The minimum condition number was obtained for σ_f around 1.00-1.25. This result is not intuitively obvious.

In any case, we infer from Table 5.4 that the convergence rate of the SIP solver behaved like:

$$\boxed{r \simeq s \cdot \frac{\pi}{\sqrt{n_1^2+n_2^2+n_3^2}}} \qquad (5.111)$$

where n_i is the unidirectional size of the grid, and s is a slowly variable function of σ_f and n_i. According to the last row of Table 5.4, the coefficient s was roughly on the order of unity.

Figures (5.8) and (5.9) display the actual sequence of residual errors obtained during the iterative solution process for the largest and smallest problems A and D listed in Table 5.4. Figure 5.8a in particular gives the Euclidean norm of the residual error $\|h^{m+1}-h^m\|$ versus the number of iterations (m) on a semi-log plot for the "1 million node" problem A. The three subproblems $\sigma_f \simeq 1.0$, 1.7 and 2.3 were solved sequentially

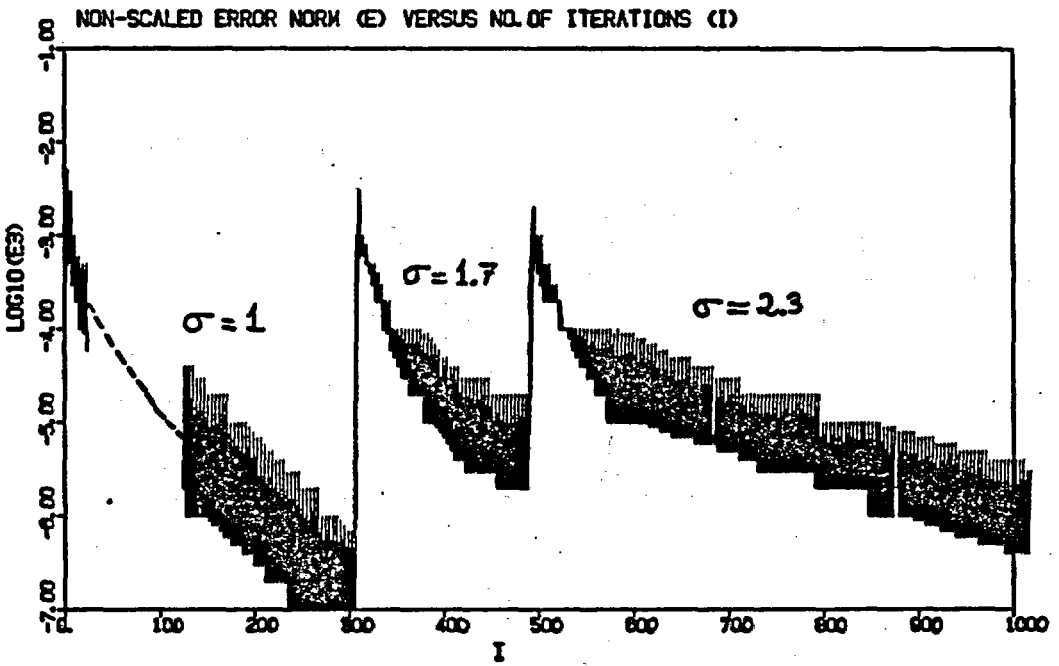


Figure 5.8 (a) Euclidean norm of the residual error versus number of iterations on a semi-log plot for problem A (1 Million nodes). The three suproblems σ=1,1.7, 2.3 were solved sequentially on a Cray 2 computer

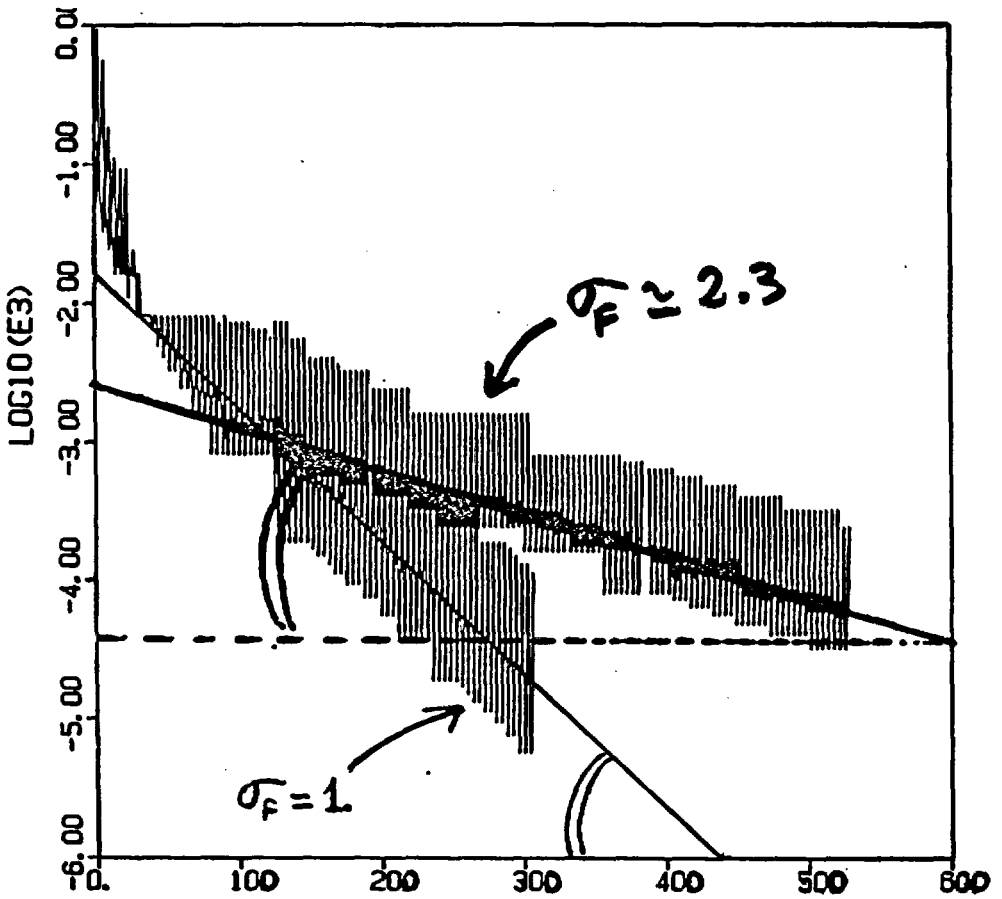


Figure 5.8 (b) Comparison of the asymptotic convergence rates for σ=1 and σ=2.3 of the 1 Million node problem A: same as Figure (5.8a) except that the residual errors have been scaled

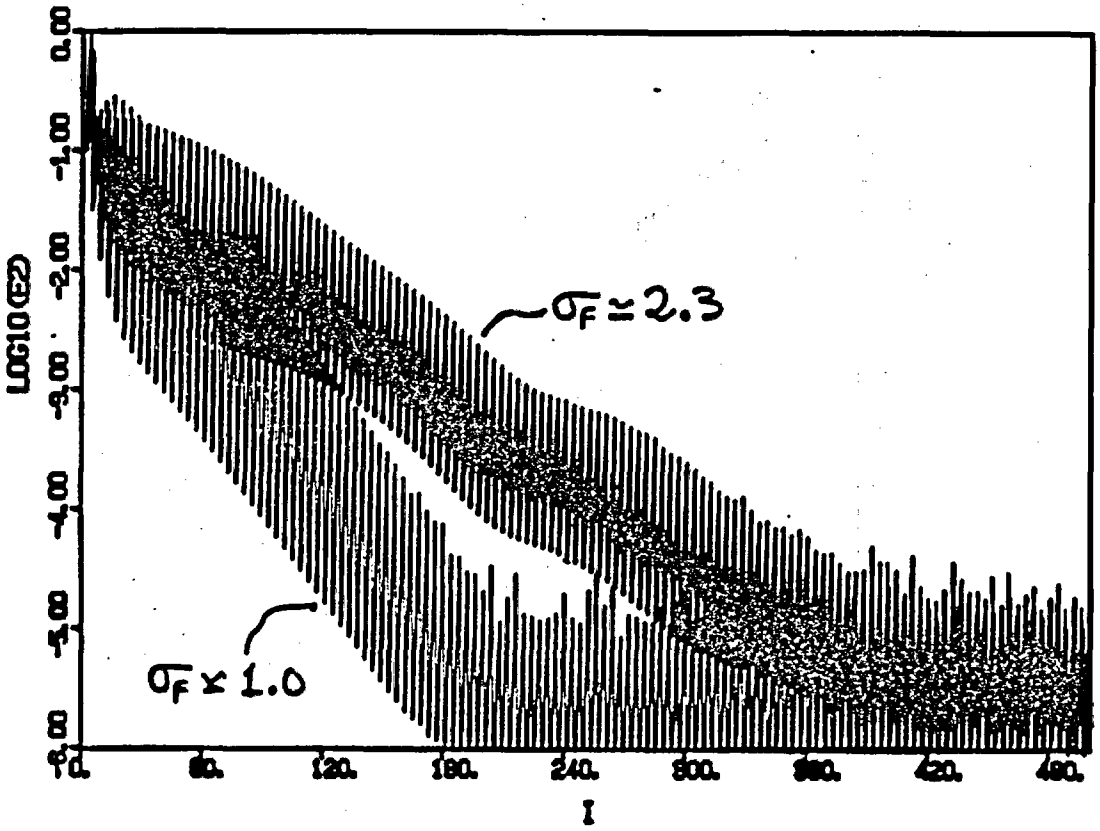


Figure 5.9 (a) Euclidean norm of the scaled residual error versus number of iterations for problem D, on a semi-log plot. The two subproblems σ=1, σ=2.3 were solved separately on a Microvax.

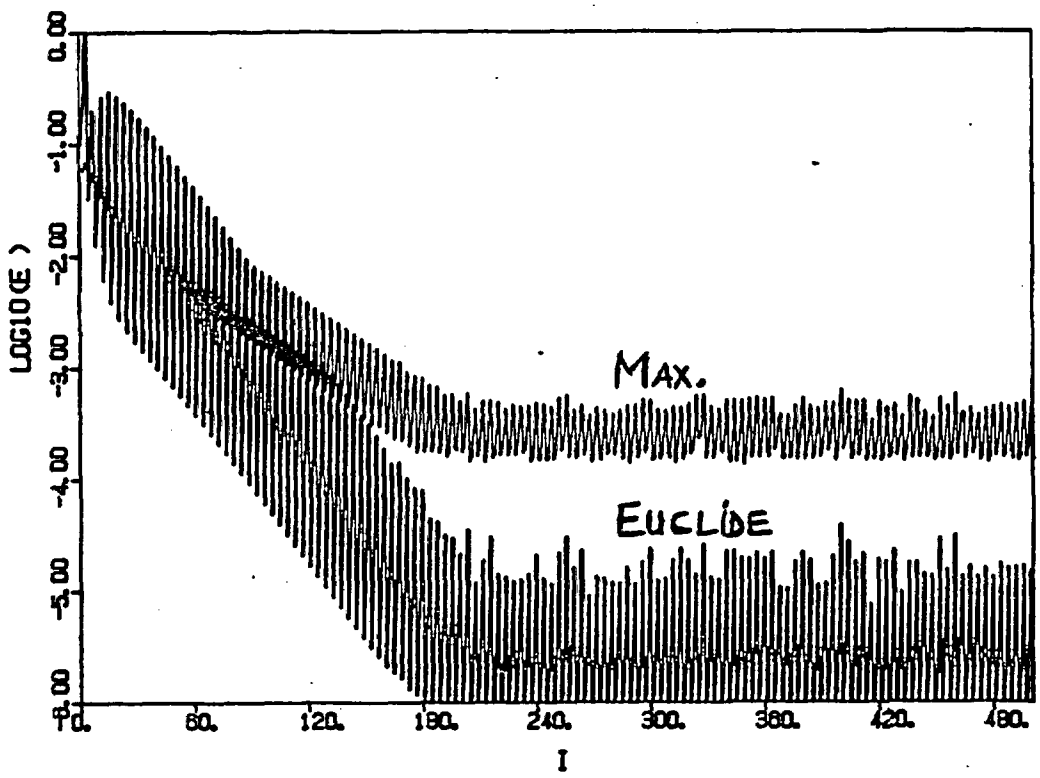


Figure 5.9 (b) Comparison of the Euclidean norm and absolute maximum norm of the scaled residual error for problem D with $\sigma_f = 1.0$.

on a Cray 2 machine, by using the last iterate of a subproblem as an initial guess for the next one. This procedure bears some resemblance with the so-called "continuation methods" to solve systems of equations whose solution depends continuously on a parameter σ: see Ortega and Rheinboldt (1970) in the context of nonlinear systems. Here the system is linear, but "difficult" to solve unless $\sigma = 0$; the three subproblems $\sigma \simeq 1, 1.7, 2.3$ correspond to three "iterates" of the continuation method.

The total number of SIP iterations required to solve accurately the three subproblems of Figure 5.8a was about 1000 iterations, which consumed a total of about 4 CPU hours on the four-quadrant Cray 2 machine of the Minnesota Supercomputer Center running in "single precision" (64-bit words).

It should be noted that the Fortran code ("Bigflo") did not fully vectorize due to the nonlinear recursions of the SIP factorization algorithm, and the backward substitution of the solution algorithm. In addition, a "slow" Fortran compiler ("cft77") was used for technical reasons. As a result, the speed-up ratio between the Cray 2 and a Microvax 2 machine was moderate, about 1 CPU hour/1 CPU minute. The CPU time was found to follow the general relation:

$$T \simeq (c_0 + c_1 \cdot \frac{m}{100}) \cdot \frac{N}{10^6} \qquad (5.112)$$

where $c_0 \simeq 5$ mn (overhead per million nodes) and $c_1 \simeq 23$ mn (per hundred iterations/per million nodes) for the Cray 2 machine. Recall that the number of iterations (m) was found to be roughly proportional to the largest unidirectional size of the grid. Using equations (5.108) and (5.111) and ignoring the overhead time gives approximately the CPU time required to decrease the residual error by 1 order of magnitude, as a function of the grid size:

$$T_1 \simeq c_1 \cdot \prod_{i=1,2,3} \left(\frac{n_i}{100}\right) \cdot \max_{i=1,2,3} \left\{\frac{n_i}{100}\right\} \qquad (5.113)$$

For a grid with equal size n in all directions, this gives simply:

$$T_1 \simeq c_1 \left(\frac{n}{100}\right)^4 \qquad (5.114)$$

For the 1 million node grid (n = 100) this yields just 23 CPU minutes on the Cray 2. However, doubling the unidirectional size of the grid (n = 200) yields an 8 million node grid, which would require about 6 CPU hours of Cray 2 time in order to decrease the residual error by just one order of magnitude! This indicates that flow problems on the order of ten million grid points or more are presently very costly or infeasible in any "reasonable" amount of time with the SIP solver. Presumably, this assessment

also holds for most of the iterative solvers discussed in our literature review.

For the flow problems reviewed in Table 5.4, fairly accurate solutions were obtained after the initial residual error was decreased by two to three orders of magnitude. The iterations were usually stopped after the relative error (defined below) reached about 10^{-2} - 10^{-3} or less. The relative solution error was evaluated a posteriori in two steps. First, the residual errors (Figure 5.8a) were normalized by a typical head variation (σ_H) which was taken to be the standard deviation of head given by the approximate spectral solutions of Chapter 3. Figure 5.8b shows the sequence of *scaled* residual errors for the 1 million node problem A, with $\sigma_f = 1$ and $\sigma_f = 2.3$. A similar graph is depicted in Figure 5.9a for the smaller problem D.

The *true* solution error (normalized by σ_H) was then evaluated according to equations (5.105)-(5.106). As an example, let us focus on problem A with $\sigma_f = 2.3$ (Figure 5.8b). The scaled residual error at the last iteration was:

$$\frac{\|h^{m+1}-h^m\|}{\sigma_H} \simeq 10^{-4}.$$

However, the asymptotic convergence rate was quite small:

$$r = - \frac{d\ell n \|h^{m+1} - h^m\|}{dm} \simeq 0.0070.$$

According to equation (5.105), we evaluated the upper bound on the *true* solution error as follows:

$$\frac{\|h - h^m\|}{\sigma_H} \leq \frac{10^{-4}}{1 - e^{-0.007}} \simeq 1.5\ 10^{-2}.$$

This shows that the solution error was at most 1.5%, relative to the magnitude of fluctuations of the solution (σ_H). Although this is certainly acceptable, it should be noticed that the "true" error (upper bound) appears to be almost 2 orders of magnitude larger than the residual error. The discrepancy was less marked in the case of smaller variability ($\sigma_f = 1.0$), and the solution error was found to be only a fraction per cent relative to σ_h. Indeed, it is clear from Figures (5.8) and (5.9) that convergence was faster in the case of smaller variability ($\sigma_f = 1$, compared to $\sigma_f = 2.3$).

We conclude that the solutions obtained with the SIP solver were highly accurate, since the solution error at the last iterate was found to be at most on the order of one percent the standard deviation of the (random) head solution itself. In our view, this is a remarkable result given the large size and high variability of the problems considered for solution (especially problem A, with 1 million nodes and σ_f=2.3).

On the other hand, we have seen by extrapolating a relation on the number of iterations versus grid size, that the solution of larger random flow problems on the order of 10 million nodes or more may not be at hand. Even if convergence could be achieved, we predict that the solution of such problems could require 1 CPU day or more on recent supercomputers like the Cray 2. Furthermore, the solution process could be overwhelmed by the accumulation of *round-off errors* before reaching a satisfactory solution error. This possibility seems to be indicated by Figure 5.9b, which shows that the residual error cannot be decreased beyond a certain level. Note that Figure 5.9b is for the "small" problem D, solved on the Microvax 2 minicomputer with 32-bit words. The accumulation of round-off errors was not observed with the Cray 2 simulations, due to the higher precision on this machine (64-bit words).

Incidentally, Figure 5.9b also compares the Euclidean norm and the absolute maximum norm of the residual error. As one could expect, using the absolute maximum norm gives a more pessimistic picture of the performance of the solver. Nevertheless, we emphasize the fact that, for all problems listed in Table 5.4, the SIP solver did converge in terms of the absolute maximum norm as well, and the *maximum* head error over the grid was only a fraction of the head standard deviation in all cases. In summary, the solution error due to the approximate

matrix solver SIP was quite small in all cases considered for simulation, both in the Euclidean norm and the absolute maximum norm.

Finally, it should be kept in mind that truncation errors will necessarily compound with the matrix solution errors to yield, presumably, a larger total error. Considering the results above and those of Section 5.2 on truncation errors, it appears that the total root-mean-square error on the hydraulic head will be at most about 5% relative to the head standard deviation. Note that this evaluation is based on the one million node problem (A) of Table 5.4 with $\sigma_f = 2.3$. The error so estimated refers of course to the exact solution of the finite-domain single-realization problem, not to the hypothetical ergodic solution of the infinite-domain problem as assumed for instance in the spectral theory.

5.4 Development and Analysis of the Nonlinear Iterative Solver for Transient Unsaturated Flow

5.4.1 Nonlinear SIP solver and nested Picard iterations:

[a] Overview of Numerical Issues:

In the case of unsaturated or partially saturated flow -- as opposed to saturated flow -- a new solution strategy must be devised in order to take into account the highly *nonlinear* nature of the algebraic finite difference system. We have chosen in this work to focus our analysis on the case of *transient* flow problems, such as local infiltration in semi-infinite unsaturated media. Steady state solutions, when they exist, can be obtained by running the transient unsaturated flow simulator for large times. An example of this can be found in Chapter 7, where large scale unsaturated flow solutions are presented for both transient and steady state cases.

In the transient regime, and when the time steps are sufficiently small, the algebraic system to be solved becomes much better conditioned than its steady state counterpart (see equation 5.31). As a consequence, the SIP matrix solver is expected to converge much faster for each time step of a transient flow problem than it does for the single step of a steady flow problem. On the other hand, the nonlinearity of the

unsaturated flow system requires prior linearization in order to obtain a solvable matrix system, presumably with iterative corrections to converge to the solution of the *nonlinear* system. Now, it is likely that there will be some restriction on the time step size in order to ensure the convergence of the iterative linearization scheme. Therefore, we expect that the interplay between time step size and convergence rate of the linearization scheme will play a major role in determining the overall efficiency of the unsaturated flow simulator (perhaps more important than the rate of convergence of the SIP matrix solver). With this "warning" in mind, we proceed to describe the actual procedure used to solve the nonlinear transient flow problem.

[b] Description of the Nonlinear-SIP Solver:

The procedure which we have developed is based on a doubly-iterative Picard scheme for solving the nonlinear finite difference system at each time step. The outer Picard iteration loop is a simple iterative predictor-corrector scheme which breaks down the nonlinear system into a sequence of linear systems. This was described in Section 5.1.3, where we developed in detail the nonlinear space-time finite difference system and its linearized version (see in particular equations 5.28-5.31). The inner Picard iteration loop corresponds to the solution of

the linearized matrix system by the SIP method, for each step of the outer iteration loop. For simplicity, the whole solution procedure will be designated as "nonlinear SIP".

The linearized system (equation 5.31) can be written for each time step ($n \rightarrow n+1$) and each outer iteration ($k \rightarrow k+1$) as:

$$A^{n+1,k} \cdot \delta h^{n+1,k+1} = \eta \cdot r^{n+1,k} \qquad (5.115)$$

where η is a new nonlinear relaxation parameter, and r is the residual for the outer Picard iteration loop ($r^k \rightarrow 0$ as $k \rightarrow \infty$). Note that the linearized vector of boundary conditions was absorbed in the residual. The numerical code accommodates boundary conditions of fixed pressure, fixed flux, or zero pressure gradient, for each of the nodes belonging to the boundary.

The linearized coefficient matrix A has the same sparse structure as the matrix previously obtained for the linear system of saturated flow -- a major advantage of the Picard linearization scheme over Newton-Raphson. More precisely, we have shown earlier (equation 5.31) that A takes the form:

$$\underline{\underline{A}}^{n+1,k} = \frac{C^{n+1,k}}{\Delta t_n} \cdot \underline{\underline{I}} + \underline{\underline{K}}^{n+1,k} \qquad (5.116)$$

where C is the linearized storage term (for transient flow only), $\underline{\underline{I}}$ is the identity matrix, and $\underline{\underline{K}}$ is the linearized unsaturated conductivity matrix, formally identical to the coefficient matrix of the saturated flow system. As mentioned earlier, the condition of matrix A improves as Δt_n decreases, due to enhanced diagonal dominance in that case.

Plugging the standard SIP iteration scheme of equation (5.81) into the linear system (5.115) yields a doubly-iterative Picard scheme of the form:

$$\boxed{\begin{aligned} y^{k+1,m} &= h^{k+1,m} - h^{k} = \delta h^{k+1,m} \\ y^{k+1,m+1} &= h^{k+1,m+1} - h^{k} = \delta h^{k+1,m+1} \\ \delta y^{k+1,m+1} &= y^{k+1,m+1} - y^{k+1,m} \\ L^{k}_{m+1} \cdot U^{k}_{m+1} \; \delta y^{k+1,m+1} &= \omega \cdot (\eta \cdot r^{k} - A^{k} y^{k+1,m}) \end{aligned}} \qquad (5.117)$$

where the time index $(n+1)$ has been dropped for clarity of

exposition. The double iteration loop (5.117) runs over $k = 1,\cdots,K$ for outer iterations, and $m = 1,\cdots,M$ for inner iterations of the SIP solver (the LU matrices vary with m because of the cyclic SIP iteration parameter τ_m of equation 5.86).

Now, for any given time step $(n \rightarrow n+1)$, let h^o be the known pressure solution at time t_n. Then, the solution at the next time step $t_{n+1} = t_n + \Delta t_n$ is obtained as follows:

(0) Define the initial guess for outer iterations:
$h^o = h(t_n), \quad k = 0$.

(1) Increment the outer iteration loop $k \rightarrow k + 1$
(Picard linearization scheme)

(2) Update nonlinear coefficients of matrices L^k, U^k, A^k and vectors r^k, b^k (all functions of h^k, known from previous outer iteration step)

(3) Define the initial guess for inner iterations:
$y^{k+1,0} = h^{k+1,0} - h^k = 0, \quad m = 0.$

(4) Increment the inner iteration loop $m \rightarrow m + 1$ (Picard iterations of the SIP solver)

(5) Compute the LU factorization and solve the system (5.117) by forward-backward substitution (SIP) to obtain $\delta y^{k+1,m+1}$

(6) Update $y^{k+1,m+1} = y^{k+1,m} + \delta y$, and iterate to step (4) unless $m = M$ or $\|\delta y\| \leq \epsilon_M$

(7) Update $h^{k+1} = h^k + y$, and iterate to step (2) unless $k = K$ or $\|y\| = \|\delta h\| \leq \epsilon_K$.

(8) The last computed vector h is the desired solution $h(t_{n+1})$ at the new time; the difference $(h_{n+1}-h_n)$ is used to compute the next time step before incrementing the time loop (not shown here).

This algorithm defines, in an extremely condensed form, the nonlinear SIP solver which forms the backbone of the flow simulator. The actual Fortran code ("Bigflo") comprises over ten thousand lines of instructions and comments. A summary narrative of this code is given in Appendix 5.D. It should be emphasized that a single program was developed to solve both steady and transient, saturated and unsaturated, deterministic and random

flow problems. These various options are accommodated by certain parameters that act as switches (transient/steady, saturated/unsaturated, etc.). The "nonlinear SIP" algorithm described above includes all the cases just considered. For instance, in the case of steady saturated flow, the time loop and the outer iteration loop are simply bypassed, or incremented only once.

Another important feature of the numerical flow simulator is its modularity. For instance, the SIP solver intervening in Step (5) is the same for saturated or unsaturated flow. This solver, and many other parts of the code, are isolated in subroutines. It may be of interest to note that the SIP solver subroutine, although quite complex, takes only a very small fraction of the whole Fortran code. However, we have found that most of the computational work was usually consumed in that small part of the object code (typically 80%-90% of the total CPU time, as estimated by a Cray software called "flowtrace").

Let us now consider how the nested inner/outer Picard iteration loops can be controlled to optimize the nonlinear solution process. The control parameters that remain to be determined for a given flow problem and a given mesh size are the following:

- Time step size (Δt_n)
- Maximum number of iterations for the inner loop (M) and outer loop (K)
- Maximum residual error for the inner loop (ϵ_M) and outer loop (ϵ_K)
- Choice of a norm for the residual error (Euclidean norm? Maximum absolute norm?)
- Relaxation parameters for the inner iterations (ω) and outer iterations (η).

These control parameters were determined empirically on a case-by-case basis, i.e., by numerical experimentation, for a number of transient test problems of two and three-dimensional infiltration in homogeneous and heterogeneous soils, with grid sizes ranging from a few thousand to a few hundred thousand nodes. The variable time-step size was controlled automatically by following the evolution of the solution, as will be explained in a later subsection. It was found that the time-step size had to be quite small in order to ensure the convergence of the outer Picard iterations (although not as small as would be needed in the case of an explicit time discretization scheme). The relaxation parameters ω and η were usually taken equal to one. Note that the inner iteration loop (SIP) was *not underrelaxed* because most cases of divergence seemed to have been

triggered by the growth of the outer iteration residuals. When divergence occurred, the simulations were simply resumed with smaller time steps, while the relaxation parameters were usually kept the same ($\omega=\eta=1$). On the other hand, recall that for linear problems of steady saturated flow, the choice of the SIP relaxation parameter was critical, and $\omega < 1$ was frequently required for SIP convergence. It appears that for highly nonlinear transient flow problems, the time step size is more important as a control parameter.

The remaining parameter (K, ϵ_K) and (M, ϵ_M) were used to control the number of iterations for the outer loop and inner loop, respectively. Typically, the *maximum* length allowed for each iteration loop was set for K=M=50. The *actual* length of each loop was controlled by the tolerances ϵ_K and ϵ_M for the outer and inner residual errors expressed in terms of pressure heads. The norm chosen for comparison was the absolute maximum of the residual errors over the grid (rather than the Euclidean norm used for steady saturated flow). The chosen tolerance was typically $\epsilon_K \simeq 0.1$ cm for the outer iterations, and $\epsilon_M = 0.01$ cm for the inner (SIP) iterations. This resulted in a very short iteration loop for the SIP solver (on the order of 1-10 iterations). The outer iteration loop was somewhat larger (1-20 iterations) depending on the time step size. In one example, it was found that a moderate increase in the time step

size resulted in an increase of the average number of iterations, in such a way that the total computational work was unchanged. However, divergence eventually occurred if the time step was increased by a larger amount. Thus, the simulations were successful only in those cases where the time step was small enough that the lengths of the inner and outer loops were kept small, say, no more than 10 inner iterations and 20 outer iterations on average.

For clarity, it may be useful to define more explicitly our concept of a "successful" simulation. First, the inner and outer residuals must both decrease monotonously on average. And second, the number of iterations for each loop must remain below the preset maximum, so that the residual at the last iteration be smaller than the preset tolerance (at least most of the time). Typically, a "successful" simulation of infiltration in dry heterogeneous soils resulted in a total of 50 iterations of the SIP matrix solver per time step (say 5 inner iterations and 10 outer iterations on average). Thus, the solution of a nonlinear flow problem over, say 100 time steps, typically required 5000 SIP iterations. This is more than would be required for the solution of a steady state, saturated flow problem (Section 5.3). The increased computational work for unsaturated flow is due to the highly nonlinear nature of the governing equation. Of course, we expect that the solution of steady state unsaturated

flow problems will be even more demanding, unless a good initial guess can be found.

[c] Brief Literature Review and Discussion:

The nonlinear solution procedure just described bears some resemblance with a number of methods proposed in the literature. In particular, various versions of nonlinear SIP solvers that are similar in some respects to the present method were developed by Trescott and Larson (1977), Kuiper (1981), and Kuiper (1987) for the simulation of groundwater flow with a variable watertable. In the first two papers, the standard SIP solver was implemented in such a way that the nonlinear coefficients were updated at every SIP-iteration. This is nearly equivalent to reducing the inner iteration loop of our solver to just one iteration (M=1). The third paper (Kuiper 1987) was devoted to the comparison of a number of variants of the SIP and ICCG solvers in conjunction with various strategies for the outer linearization loop, including SIP-Picard and SIP-Newton strategies with only 1-5 inner iterations of the SIP solver. Finally, Cooley (1983) addressed the problem of partially saturated flow with seepage faces, using a complex procedure be described as a combination of Newton-Raphson iteration, successive approximation, and (SIP).

It should be noted that the numerical experiments reported by these authors involved rather small, mildly nonlinear problems, with grid sizes generally below 1000 nodes. Unfortunately, we have found it quite difficult to draw conclusions from their work regarding the optimal design of the nonlinear solvers. The nonlinear systems to be solved are sometimes so complex that they could be sensitive to even minute details of implementation. One single fact seems to emerge however: most authors have chosen a solution strategy that imposes a very small number of matrix solver iterations between each nonlinear coefficient updates. This is similar to what occurred in actual practice with the more flexible nonlinear-SIP method developed in this work.

More details on certain aspects of the unsaturated flow simulator will be given below, particularly concerning the dynamic control of time step size and domain size (in cases where a variable domain evolving with the solution makes sense), as well as mass balance computation, and other related issues. A semi-empirical analysis of space-time resolution requirements will also be developed in order to obtain heuristic criteria for the choice of mesh size and time step size, particularly for ensuring the convergence of the linearization scheme. In addition, numerical experiments for a variety of test problems will be presented in order to explore the actual capabilities of

the unsaturated flow simulator. However, the analysis and physical interpretations of the numerical solutions obtained for *random* unsaturated flow problems is postponed to Chapter 7. It should be noted that the largest such problem analyzed in this work involved a three-dimensional grid size on the order of 0.3 million nodes, and over a hundred steps in time. As far as we know, this problem size is 1 to 2 orders of magnitude larger than currently available simulations of unsaturated flow published in the literature. Even for modest size problems, it does not seem that the degree of variability considered in this work, with a node-by-node variation of the random constitutive properties of unsaturated porous media, has ever been considered elsewhere for direct numerical simulations. The present flow simulator appears therefore as a unique "high resolution" tool for exploring highly heterogeneous nonlinear unsaturated flow phenomena.

5.4.2 Truncation errors, nonlinear stability, and space-time resolution requirements

[a] Methodology

In this subsection, we attempt by various methods to evaluate the numerical requirements for convergence and accuracy of the unsaturated flow simulator. The major numerical

difficulty appears to be the highly nonlinear nature of the governing flow equation, with the soil moisture capacity and unsaturated conductivity curves given by equations (5.19) and (5.22). In the case of transient infiltration in dry soils, we have observed that the solution could diverge after just one or a few time steps if the initial time step size was taken too large. The problem becomes naturally more severe in the case of spatially variable soils: instabilities can be triggered at any time as a moving infiltration front encounters zones of higher or lower conductive properties. In this light, it seems worthwhile to look for constraints on the mesh size and time step size that will guarantee the convergence of the Picard linearization scheme (outer iteration loop of the nonlinear-SIP solver). The question of accuracy of the finite difference approximation can also be examined in terms of truncation errors. It seems however futile to draw conclusions from truncation analysis without taking into account the errors due to *linearization*. This difficult enterprise was not pursued in this work. Instead, the issues of accuracy (truncation error) and convergence (nonlinear stability analysis) will be examined separately. The latter view-point will lead to some specific numerical requirements, however without rigorous proof.

[b] Truncation Error Analysis:

Let us develop a simplified truncation error analysis for the one-dimensional, transient unsaturated flow equation with spatially constant soil properties. Note that we do not attempt to develop closed form expressions for the solution error as was done in Section 5.2 for the case of random saturated flow. Briefly, the truncation error is defined by the expression:

$$T(h_i) = \hat{L}(h_i) - L(h_i). \qquad (5.118)$$

where $L(h)$ represents the (vanishing) operator corresponding to the one-dimensional transient unsaturated flow equation:

$$L(h) = C(h) \frac{\partial h}{\partial t} + \frac{\partial}{\partial x} (K(h) \cdot (\frac{\partial h}{\partial x} + g)) = 0 \qquad (5.119)$$

Note that g indicates the acceleration of gravity: take $g = 0$ for horizontal flow, and $g = +1$ for vertical flow with the x-axis upwards. On the other hand, $\hat{L}(h_i)$ represents the finite difference operator defined analogously to equations (5.24)-(5.30). In this discrete operator, the midnodal unsaturated conductivity $K(h_{i+\frac{1}{2}})$ is evaluated by the geometric mean $\hat{K}_{i+\frac{1}{2}}$, as in equation (5.29).

The truncation error (5.118) can be expressed formally

by using Taylor developments analogous to those of Section 5.2 (e.g., equation 5.36). The calculations are tedious but straightforward, and its details will not be reproduced here (some intermediate results can be found in Vauchin et al., 1979, p. 40). When the truncation error due to the approximate evaluation of mid-nodal conductivities ($\hat{K}_{i+½}$) is ignored, the resulting truncation error (T_i^n) at the nodes of the space-time grid takes the form:

$$T(h_i^n) = \frac{\Delta x^2}{24} \cdot \left\{ \frac{\partial}{\partial x} \left[K \frac{\partial^3 h}{\partial x^3} \right] + \frac{\partial^3}{\partial x^3} \left[K \left[\frac{\partial h}{\partial x} + g \right] \right] \right\}_i^n$$

$$+ \Delta t \cdot \left\{ \frac{\partial}{\partial x} \left[K \frac{\partial}{\partial t} \left[\frac{\partial h}{\partial x} \right] \right] - \frac{1}{2} C \frac{\partial^2 h}{\partial t^2} \right\}_i^n \qquad (5.120)$$

On the other hand, when $\hat{K}_{i+½}$ is evaluated by the geometric mean weighting scheme, the error ($\hat{K}$-K) needs also to be taken into account. Assuming an exponential conductivity - pressure relation:

$$K(h) = K_s \exp(\alpha \cdot h) \qquad (5.121)$$

where both K_s and α are assumed constant, the mid-nodal conductivity error takes the form:

$$\hat{K} - K = \frac{\Delta x^2}{8} \cdot \alpha K \frac{\partial^2 h}{\partial x^2} . \qquad (5.122)$$

When this discrepancy is taken into account in the Taylor developments, an additional term is obtained in equation (5.120). This new term takes the form:

$$+ \frac{\Delta x^2}{8} \cdot \frac{\partial}{\partial x} \left[\alpha \frac{\partial^2 h}{\partial x^2} \cdot K \left[\frac{\partial h}{\partial x} + g \right] \right] \qquad (5.123)$$

and the total truncation error results from adding (5.123) to the right-hand side of (5.120).

A few remarks can be made about equations (5.120)-(5.123). First of all, we find that the order of accuracy of the finite difference approximation is $O(\Delta x)^2$ in space and $O(\Delta t)$ in time. However, we expect a degradation of the spatial accuracy when the coefficients are random fields instead of constants (by analogy with the results of Section 5.2). Our second remark is that the accuracy in time depends on the rate of change of the pressure gradient and on the second derivative $\partial^2 h/\partial t^2$. Finally, it is interesting to note that the error due to inaccurate weighting of mid-nodal conductivities (equation 5.123) is roughly proportional to $\partial^3 h/\partial x^3$ (this is in fact exactly true in the steady state case of 1 dimension). Thus, the error is largest in regions of rapid changes of the

curvature of h(x), which are located just above and just below the inflexion point of the wetting front.

[c] Nonlinear Stability Analysis:

Unfortunately, it appears that our formal evaluation of truncation errors does not lead to any useful estimates of space-time resolution requirements. This is due to the fact that the approximate, iterative linearization of the finite difference system was left out of the analysis. Here, we propose an indirect way of including the "linearization error" by considering only one iteration of the outer loop of the nonlinear solver. For one-dimensional flow, the corresponding finite difference system can be expressed as:

$$C_i^n \frac{h_i^{n+1}-h_i^n}{\Delta t_n} = K_{i+\frac{1}{2}}^n \left[\frac{h_{i+1}^{n+1}-h_i^{n+1}}{\Delta x}\right] - K_{i-\frac{1}{2}}^n \left[\frac{h_i^{n+1}-h_{i-1}^{n+1}}{\Delta x}\right] \tag{5.124}$$

$$+ g \cdot \frac{K_{i+\frac{1}{2}}^n - K_{i-\frac{1}{2}}^n}{\Delta x}$$

where n indicates the time level, and g is zero for horizontal flow, one for vertical flow. This formulation reveals that the nonlinear gravity term, containing g, is in fact treated *explicitly* during the first outer iteration of the nonlinear solver, even though we used a so-called fully implicit

time discretization scheme. Let us now examine how this could affect the stability of the numerical solution.

One possible way of investigating the specific effects of linearization of the gravitary term is to develop a Fourier stability analysis of equation (5.124) with "frozen coefficients" on the storage term and diffusive term, while the nonlinearity of the gravity term is explicitly taken into account. In the case of the exponential conductivity model (5.121), the gravity term takes the special form:

$$g \cdot \frac{K^n_{i+\frac{1}{2}} - K^n_{i-\frac{1}{2}}}{\Delta x} = g \cdot \frac{K^n_i}{\Delta x} \cdot \left\{ \exp\left[\alpha \frac{h^n_{i+1} - h^n_i}{2}\right] - \exp\left[\alpha \frac{h^n_{i-1} - h^n_i}{2}\right] \right\}$$

which may be approximated as:

$$g \frac{K^n_i}{\Delta x} \frac{\alpha}{2} (h^n_{i+1} - h^n_{i-1})$$

provided that $\alpha(h_{i+1} - h_i)$ be on the order of unity or less. Thus, we find that the FD scheme (5.124) is approximately equivalent to:

$$-R_{i-\frac{1}{2}} \cdot h^{n+1}_{i-1} + (1 + 2\bar{R}_i) h^{n+1}_i - R_{i+\frac{1}{2}} h^{n+1}_{i+1} \simeq$$

$$- \frac{1}{2} g \alpha \Delta x R_i \cdot h^n_{i-1} + h^n_i + \frac{1}{2} g \alpha \Delta x R_i h^n_{i+1} \qquad (5.125)$$

where R is a dimensionless number proportional to $\Delta t/\Delta x^2$:

$$R_{i(\pm\frac{1}{2})} = \frac{K_{i(\pm\frac{1}{2})}}{C_i} \cdot \frac{\Delta t}{\Delta x^2} \; ; \; \overline{R}_i = \frac{R_{i-\frac{1}{2}}+R_{i+\frac{1}{2}}}{2} \; .$$

Now, a standard Fourier stability analysis of equation (5.125) with the R coefficients "frozen" (see for instance Ames, 1977) yields the complex time amplification factor of the Fourier modes of h(x):

$$\rho \simeq \frac{1+j\cdot g\ \alpha\Delta x\ R_i \sin\ k\Delta x}{1+(R_{i+\frac{1}{2}}+R_{i-\frac{1}{2}})(1-\cos\ k\Delta x) - j\cdot(R_{i+\frac{1}{2}}-R_{i-\frac{1}{2}})\sin\ k\Delta x}$$

where k is the discrete wavenumber taking values $(\pi/L, \cdots, n\pi/L)$. After some manipulations, the square-modulus of this amplification factor takes the form:

$$|\rho|^2 \simeq \frac{1 + [g\ \alpha\Delta x\ R_i\ \sin\ k\Delta x]^2}{[1+(R_{i+\frac{1}{2}}+R_{i-\frac{1}{2}})(1-\cos\ k\Delta x)]^2 + [\ (R_{i+\frac{1}{2}}-R_{i-\frac{1}{2}})]\sin\ k\Delta x]^2} \qquad (5.126)$$

Clearly, when the gravitary term disappears (horizontal flow: g = 0) the amplification factor is always less than one, and the FD scheme is then inconditionally stable. However, in the presence of the gravity term (vertical flow: g = 1) equation

(5.126) shows that a constraint on the mesh size will be necessary in order to guarantee stability.

The exact stability condition for the case $g \neq 0$ can be found by requiring that the denominator be larger than the numerator in (5.126). To simplify the analysis, let us assume $R_i \simeq R_{i+\frac{1}{2}} \simeq R_{i-\frac{1}{2}}$. The stability condition is then:

$$1 + R_i \cdot [1-(g \frac{\alpha \Delta x}{2})^2] \cdot (1-\cos\ k\Delta x) \geq 0.$$

This condition is required for all wavenumbers ($k=\pi/L, \cdots, n\pi/L$) where n is the unidimensional size of the grid. Since $R_i > 0$, this condition is always satisfied in the case $g = 0$, as expected. For non-horizontal flow ($g \neq 0$), it is easily seen that the inequality above will be satisfied if and only if:

$$(g \frac{\alpha \Delta x}{2})^2 \leq 1 + (2R_i)^{-1}.$$

Plugging $g = 1$ for vertical flow, and using the previous definition of R_i give finally the "nonlinear stability condition":

$$\boxed{\alpha\ \Delta x \leq 2 \cdot \sqrt{1+(2 \frac{K_i}{Ci} \frac{\Delta t}{\Delta x^2})^{-1}}.} \qquad (5.127)$$

The second term inside the square root corresponds to the well known stability condition for explicit time discretization schemes (whereas the discretization used in this work is of implicit type). If $\Delta t/\Delta x^2$ is taken to be significantly larger than the inverse diffusivity C/K, the condition (5.127) becomes equivalent to a constraint on the *grid Peclet number*, of the form:

$$\boxed{Pe = \alpha \Delta x \leq 2 \ .} \qquad (5.128)$$

For completeness, note that in the case where inequality (5.128) is not satisfied, then the stability condition (5.127) implies a rather stringent constraint on the time step:

$$\boxed{2 \frac{K_i}{C_i} \frac{\Delta t}{\Delta x^2} \leq \frac{1}{\frac{\alpha \Delta x}{2} - 1} \text{ if } \alpha \Delta x > 2 \quad .} \qquad (5.129)$$

In conclusion, equations (5.127)-(5.129) show that the linearized finite difference scheme used to approximate the unsaturated flow equation may not be stable unless the grid Peclet number $Pe = \alpha \Delta x$ is less than 2. In the more general case of multidimensional space, this gives a constraint on the vertical mesh size, Δx_1, of the form $\Delta x_1 \leq 2 \alpha^{-1}$.

For heterogeneous soils however, the coefficient α

will be spatially variable, and the constraint will not be satisfied at every nodes of the grid in actual practice. Therefore, attention is called on the time step constraint (5.129). This condition is rather stringent, having the same form as the condition for stability of *explicit* time integration schemes. This indicates that the best strategy might be to minimize as much as possible the vertical grid Peclet number, e.g.:

$$\boxed{Pe_i = \bar{\alpha}\Delta x_i << 2} \qquad (5.130)$$

where $\bar{\alpha}$ is some average of $\alpha(\underline{x})$ over the grid, or some mean value defined in ensemble space.

[d] Interpretation of the Peclet Constraint and Discussion:

It may be interesting to note that the inverse of $\alpha = d\ell nK/dh$ has been interpreted in the literature as a pore size distribution index, or as the typical thickness of the capillary fringe (Yeh et al., 1985). Another interesting interpretation is that α represents a "gravity/diffusion" ratio. Both types of interpretations seem qualitatively correct. Observations show that α is largest in coarse soils, where gravity effects are significant and the pore size distribution is typically quite narrow. A review of various interpretations of α in the literature can be found in a recent paper by White and Sully

(1987), who also contribute their own. It seems particularly useful to view α as a gravity/diffusion ratio in the context of numerical analysis. Ababou (1981) suggested this interpretation by rewriting the pressure-based flow equation (5.119) in terms of the "Kirchhoff potential":

$$\phi(h) = \int_{-\infty}^{h} K(h')dh' \ ,$$

named after Kirchhoff (1894) for his work on heat conduction. In the case of an exponential conductivity relation of the type (5.121), it turns out that $\phi = \alpha^{-1}K$, and the multidimensional flow equations can then be expressed by using the unsaturated conductivity as the dependent variable:

$$\frac{\partial K}{\partial t} = D \ . \ \{\nabla^2 K + \alpha \cdot (\underline{g} \cdot \underline{\nabla} K)\} \qquad (5.131)$$

where D is a nonlinear diffusivity function ($D = K/C$) and $\underline{g}$ is the gravity vector, for instance $\underline{g}=(1,0,0)$ for a 3D system of coordinate with the x_1-axis upwards. The so-called Kirchhoff equation has been used extensively in the area of soil water physics since the early works of Wooding (1968), Philip (1969), Raats (1971), and Parlange (1972) among others.

It is now clear that the coefficient α takes the form:

$$\alpha = \frac{\alpha D}{D} = \frac{dK/d\theta}{D} = \frac{V}{D}$$

where θ is the volumetric water content of the soil. This shows that α is indeed the ratio of gravitary versus diffusive coefficients of the conductivity-based (Kirchhoff) equation given above. In particular, note that αD scales like a velocity ($V=\alpha D$). This finally justifies our interpretation of $\alpha\Delta x$ as a *grid Peclet number*:

$$Pe = \alpha\Delta x = \frac{V\Delta x}{D} \quad .$$

To complete the analogy, observe that the Kirchhoff equation (5.131) is equivalent to the equation governing heat conduction in a body that moves with velocity V with respect to the heat source (see Carslaw and Jaeger, 1959, for heat conduction problems). Similarly, the K-based equation is also equivalent to the convection-diffusion equation governing solute transport in a porous medium, with diffusion coefficient D and water velocity V. In the case of unsaturated flow, D is the soil moisture diffusivity and $V=dK/d\theta$ gives in certain cases the rate of advance of the wetting front in the vertical direction. The proposed stability constraint on the grid Peclet number (equations 5.128 or 5.130) is most usually taken into account in the area of solute transport modeling, but does not seem to have been invoked in the context of unsaturated or two-phase flow

simulation, as we do here.

In summary, we have found that a constraint on the grid Peclet number $Pe = \alpha \Delta x_i$, analogous to the ratio of velocity versus diffusion coefficient in convection-diffusion problems, must be satisfied in order to ensure the stability of the finite difference approximation of the *nonlinear* unsaturated flow equation. Here, we have assigned a particular meaning to the notion of stability, such that the error due to linearizing the gravitary term of the equation is taken into account, while all other nonlinear coefficients are "frozen". In addition, our analysis was based on the assumption that only *one outer iteration* of the nonlinear solver was performed for each time step (see equation 5.124). In spite of the many approximations involved, we believe that the Peclet number constraint is meaningful and gives an approximate condition for the stability and convergence of the outer Picard iterations of the nonlinear SIP solver.

Accordingly, the Peclet number constraint (5.130) was taken into account in all of our numerical simulations. Note that the coefficient $\bar{\alpha}$ lies in the range 0.01-0.10 cm^{-1} for natural soils (clayey soils to coarse sands). According to the Peclet constraint ($\alpha \Delta x_i << 2$) the vertical mesh size should not be taken greater than 10-100 cm, depending on the type of soil.

We have found this rule to be quite useful, although far from sufficient for ensuring stability without additional constraints on the time step. Indeed, equation (5.129) suggests a more pessimistic view of the numerical requirements, particularly for heterogeneous soils where the Peclet constraint may not be enforced at all nodes of the grid.

Thus, it may be interesting to examine how instabilities could be triggered locally in regions where the Peclet constraint is not satisfied. Equation (5.126) and the inequality below specifically show that the high wavenumber Fourier nodes are the most unstable (fluctuations at the mesh scale are amplified faster). Moreover, the instability will be more severe for large values of $\Delta t/\Delta x^2$ and for large values of the local diffusivity (e.g., near saturation).

However, results from Fourier stability analysis may be too approximate to be reliable in practice. It remains unclear whether local scale instabilities, such as due to high Peclet number, will actually grow and contaminate all scales of fluctuations from mesh size to domain size. Another approach, based on the requirement that the spectral radius of the nonlinear iteration matrix be less than unity (Ortega and Rheinboldt, 1970, 7.1 and 10.1), suggested that the condition for convergence of the nonlinear iterations could be of the form:

$$Pe_m = \alpha(\underline{x})\Delta x_m \leq 2 \cdot \left|\frac{\partial H}{\partial x_m}\right|^{-1} \qquad (5.132)$$

which means that the mesh size in any direction (m = 1,2,3) must be taken small in inverse proportion of the hydraulic gradient ($\partial H/\partial x_m$) in that direction. This is obviously a more severe constraint than the standard Peclet number condition. It indicates that the presence of sharp hydraulic gradients, such as occur at a wetting front, could trigger global divergence of the nonlinear SIP solver, even when the mesh size is small enough that the Peclet condition $\alpha\Delta x << 2$ is satisfied. This observation seemed to be confirmed by numerical experiments, as most cases of divergence occurred during the early times of infiltration in very dry soils (sharp fronts).

5.4.3 Numerical experiments and test of problem solving capabilities:

In this last subsection, we present the results of a number of numerical experiments in order to test the problem solving capabilities of the unsaturated flow simulator. Accordingly, we focus have mainly on numerical issues, and only occasionally on physical interpretation. The numerical solutions obtained for large problems of three-dimensional infiltration in *random* soils will be presented and discussed more thoroughly in

Chapter 7). The order of presentation is as follows. We first discuss the methodology and present an overview of model problems for testing the unsaturated flow simulator. Some details of implementation of the code are then described, notably concerning the dynamic control of time step size, the variable domain size, and the algorithm used for mass balance computation. Numerical experiments are finally presented, for increasingly complex problems, in view of testing various features of the unsaturated flow simulator.

[a] Overview of Test Problems and Methodology:

The complexity of the random unsaturated flow problem is such that it does not appear possible to devise a unique, well-defined procedure to check the accuracy of the numerical solution, as was done earlier in the linear case (steady saturated flow, Sections 5.2 and 5.3). In the present case, the best approach seems to be a "divide and conquer" strategy, whereby several isolated features of the code are checked through a series of simple test problems. The particular tests considered in this work can be broadly classified as follows:

(i) Comparison with exact solutions:

Quasi-analytical solutions are available for instance in the case of two-dimensional infiltration from a

surface strip source in a homogeneous soil. However, the unsaturated soil properties must be of a special type, assuming in particular a constant soil moisture diffusivity. This is usually an unrealistic assumption for transient infiltration problems (Ababou, 1981). Nevertheless, since the conductivity-pressure function is exponential, the comparison of the numerical and analytical solutions provides a test of the ability of the nonlinear SIP solver to converge to the correct nonlinear solution. Another quasi-analytical solution for one-dimensional flow with constant coefficients was also used for debugging purposes: the flow simulator was run in the unsaturated mode with constant coefficients in order to test various algorithms such as the effect of variable time step size.

(ii) Self-benchmark procedures:

We refer to "self-benchmark" as a method for testing some special feature of the flow simulator in complex cases where no analytical solution is available (highly nonlinear and random soil properties). The most obvious self-benchmark of a numerical flow simulator is the computation of global *mass balance*. Our flow simulator automatically computes the mass balance by spatial integration of soil moisture and calculation of

the mass entering or leaving the system at the boundaries. Another kind of self-benchmark procedure consists in testing the sensitivity of the numerical solution with respect to mesh size and time step size, using as a reference the solution obtained with the highest space-time resolution. The obvious limitation of this kind of test is its high computational cost: thus, in actual practice, such tests were limited to fairly small domain size. A self-benchmark test was also conducted to check whether the solutions obtained with variable domain size and fixed domain size coincide. Finally, it should be kept in mind that the iterative matrix solver, which is one of the key components of the unsaturated flow code, was already tested extensively (Section 5.3). The method we used there can be viewed as a particular type of self-benchmark, conducted by running the SIP solver over a very large number of iterations and monitoring the convergence rate of residual errors, for a variety of linear flow systems. However, it does not appear feasible to test the *nonlinear* iterative solver in the same systematic way.

(iii) Qualitative tests:

Certain anomalies of the numerical solutions can

sometimes be detected by examining the numerical solutions visually, e.g., by looking at a plot of the pressure head contours, or at pressure profiles along one-dimensional transects. For example, in the case of strip source infiltration in a perfectly layered soil, one expects the solution to be perfectly symmetric about the vertical plane running through the middle of the strip. This and a number of other features based on physical principles, can be used to detect possible inaccuracies in the numerical solutions. A few simple test problems were devised for this purpose, including one-dimensional and two-dimensional infiltration in homogeneous soils (with various types of boundary conditions), and two-dimensional strip source infiltration in horizontally layered and vertically layered soil systems. In addition, some more complex infiltration problems in three-dimensional random soils are also discussed briefly (the complete analysis of these numerical solutions is postponed to Chapter 7, as mentioned earlier).

In summary, we have adopted a variety of procedures to test the problem solving capabilities and reliability of the unsaturated flow simulator, including comparisons with known analytical solutions, self-benchmark procedures, and qualitative

evaluation of the reliability of numerical solutions based on physical principles. Let us start by a description of some of the most relevant details of implementation of the flow simulator, along with some related numerical experiments.

[b] Control of Variable Time Step, Moving Boundaries, and Mass Balance:

The *variable time step size* is computed as follows. First, an initial time step is prescribed by the user (or else a very small initial time step is computed by the code). Second, the code evaluates the maximum possible value of the global spatial variation of pressure head that could occur within the flow domain (Δh_{stab}). Typically, this pressure variation is estimated by taking the difference between the largest pressure at the boundaries and the initial pressure within the flow domain, that is:

$$\Delta h_{stab} = |h_{surf} - h_{in}|$$

where h_{surf} is either the fixed pressure at the infiltration surface, or, in case or a flux condition, the pressure obtained by solving the equation $K(h) = q_{surf}$. Finally, the variable time step size Δt_n for the next time step $t_n \rightarrow t_{n+1}$ is computed by the following algorithm ($n > 1$):

$$\Delta t_n = \text{Min} \left\{ \rho . \Delta t_{n-1} \, , \; \frac{\| h^{n+1} - h^n \|_{max}}{\Delta h_{stab.}} \cdot \Delta t_1 \right\} \qquad (5.133)$$

where Δt_1 is the first (initial) time step, ρ is an amplification ratio used for limiting the rate at which the time step may grow, and the norm $\| h \|_{max}$ represents the absolute maximum of $h(\underline{x})$ over the grid. In practice, the amplification ratio ρ should be only slightly greater than unity, e.g., $\rho = 1.05$ up to $\rho = 1.25$ at most. Our experience for infiltration problems with initially dry soils indicates that equation (5.133) usually leads to a sharp increase of the time step at early times, reaching a sill after a relatively short time of infiltration.

Figures 5.10(a),(b), and (c) show the growth of the time step for three simulations of infiltration in dry soils (plotted against number of outer iterations rather than time). The first of these figures corresponds to a problem of one-dimensional infiltration under fixed pressure (h=0) in a dry sand (Dek soil) whose unsaturated properties were given earlier in Figure 5.3. The second figure corresponds to a two-dimensional infiltration in the same soil, with a fixed pressure (h=0) at the surface of a strip source. The third and last figure corresponds to a problem of two-dimensional strip source infiltration in a two-layer system of sandy soils of moderate contrast, with a relatively high flux condition

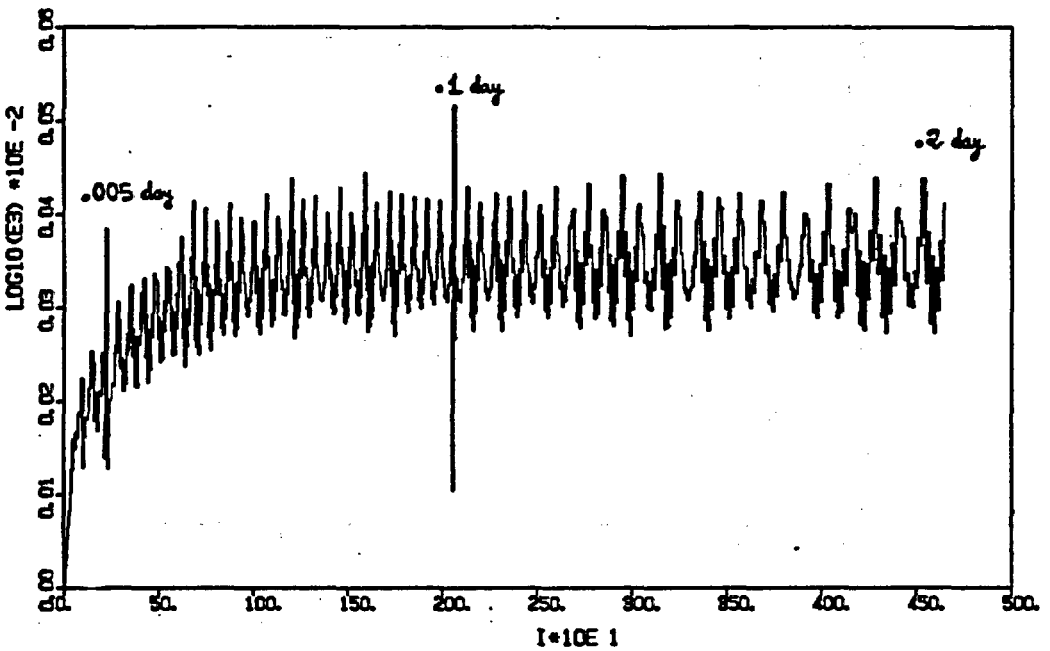


Figure 5.10 (a) Evolution of the time step size (plotted against number of outer iterations) for one-dimensional infiltration in a dry sand with fixed pressure h = 0 at soil surface.

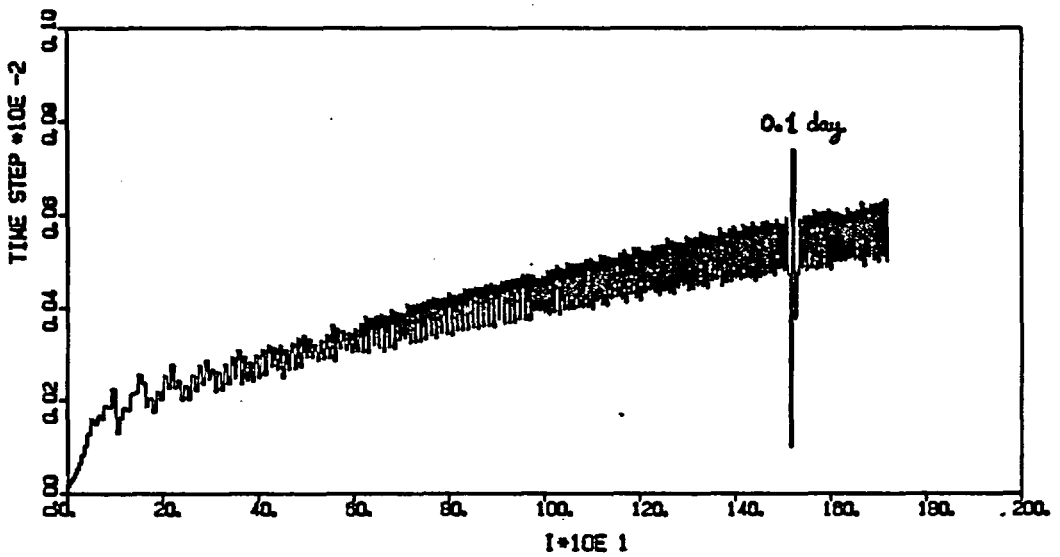


Figure 5.10 (b) Evolution of the time step size (plotted against number of outer iterations) for two-dimensional infiltration with fixed pressure h = 0 on a strip source (same soil as 5.10a).

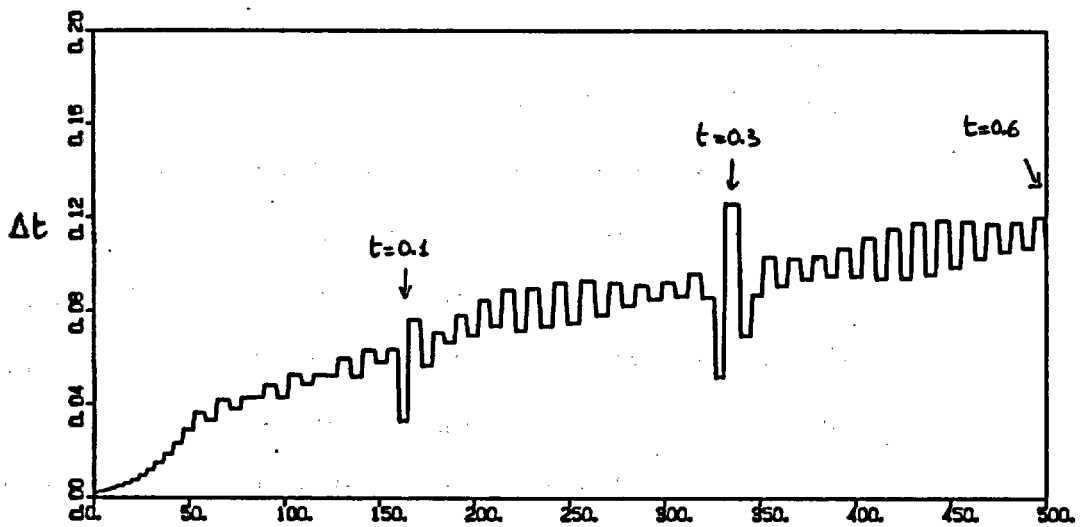


Figure 5.10 (c) Evolution of the time step size (plotted against number of outer iterations) for two-dimensional infiltration with fixed flux q = 12 cm/day on a strip source (two-layered sandy soil, top layer same as in 5.10a and b)

prescribed at the surface of the strip (q = 12 cm/day, one order of magnitude lower than the saturated conductivity of the top layer). In all cases, the initial pressure was on the order of -100 cm, corresponding to an initial conductivity on the order of 0.1 cm/day or less.

Based on these and other numerical experiments, the time-stepping algorithm appeared to be well-behaved, and did not generate instabilities. However, this algorithm was not flexible enough to handle properly large-time simulations. In such cases, the simulation was processed in several pieces: whenever the time step size appeared too small and did not increase, the simulation was stopped and resumed with a larger "initial" time step Δt_1. It is expected that a more satisfactory time-stepping algorithm could be obtained by taking into account the maximum flux over the grid (Ababou, 1981) and/or some global property of the numerical solution such as the rate of advance of the wetting front, the mean pressure gradient, or the global mass balance. Other methods of control of the time step size for similar flow problems can be found in Hanks and Bouwer (1962), Edwards (1972), Ababou (1981), and Dave and Mathis (1981). The latter authors used mass balance to adjust the time step size in their "adaptive grid" model of one-dimensional unsaturated flow.

A *variable domain size* was used for multidimensional infiltration problems on very dry soils. The size of the computational domain was controlled by moving the artificial boundaries in such a way that they always remained far from the "wetting front". The rationale behind this procedure is that, for any finite time, there exists a region beyond the wetting front where the pressure has not yet increased from its initial value. It should be noted however that this property holds only in the case of highly nonlinear coefficients and a *very dry initial state* ($h_{in} \rightarrow \infty$). The second condition is not exactly satisfied in practice. There may be a significant amount of gravity-driven flow outside the wetted zone in the case of imperfectly dry soils. For heterogeneous soils, this "natural flow" will be even more complex, with gravity acting in the vertical direction and soil moisture "diffusion" in the horizontal. Nevertheless, the variable domain size algorithm was found useful in a number of cases, leading to significant savings in computational work when the soil was dry enough that the "wetting front" could be tracked accurately.

Figure 5.11 gives a schematic representation of the "variable domain" procedure in the case of strip source infiltration in two dimensions (the three-dimensional case is treated in a similar way). The procedure can be summarized as follows. A small initial domain size must be specified by the

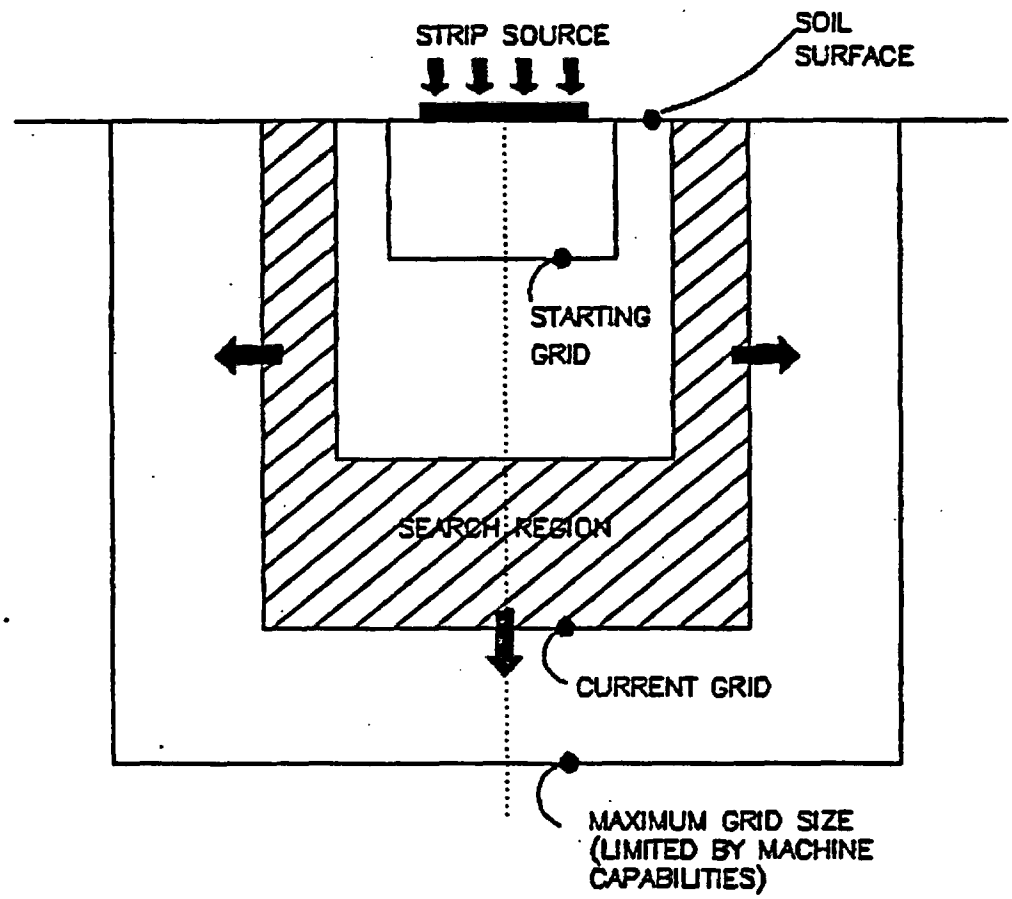


Figure 5.11 Representation of the "variable domain" procedure in the case of infiltration from a strip source. The thick arrows indicate the movement of artificial boundaries. In this example, the soil surface is the only fixed boundary.

user, along with the maximum allowable size of the domain and the definition of "artificial" (moving) boundaries, as opposed to fixed boundaries. The numerical solution at any given time step is computed based on the current domain size, with fixed pressure conditions on the artificial boundaries ($h = h_{in}$). Each artificial boundary is then moved away from the wetted zone if the maximum pressure change $\|h-h_{in}\|_{max}$ is larger than a preset tolerance (within the "search region" depicted in Figure 5.11). The algorithm is such that each boundary can move separately at its own rate, depending on the shape of the wetted zone. The displacement of a moving boundary was taken equal to the depth of the search region ($3\Delta x$). It should be noted that the grid itself was not deformed in the process, i.e., the mesh (Δx_1, Δx_2, Δx_3) remained constant in space as the size of the domain was increased. The design of a truly adaptive grid model would pose difficult problems of interpolation/extrapolation in the case of highly variable or random coefficients.

Figures (5.12) and (5.13) show two cases where the variable domain procedure worked well. Figure 5.12(a) shows the pressure head contours obtained for two-dimensional strip source infiltration on a dry sand, using the variable domain procedure with three moving boundaries. The solution obtained with a fixed domain size (Figure 5.12.b) is visually indistinguishable. A

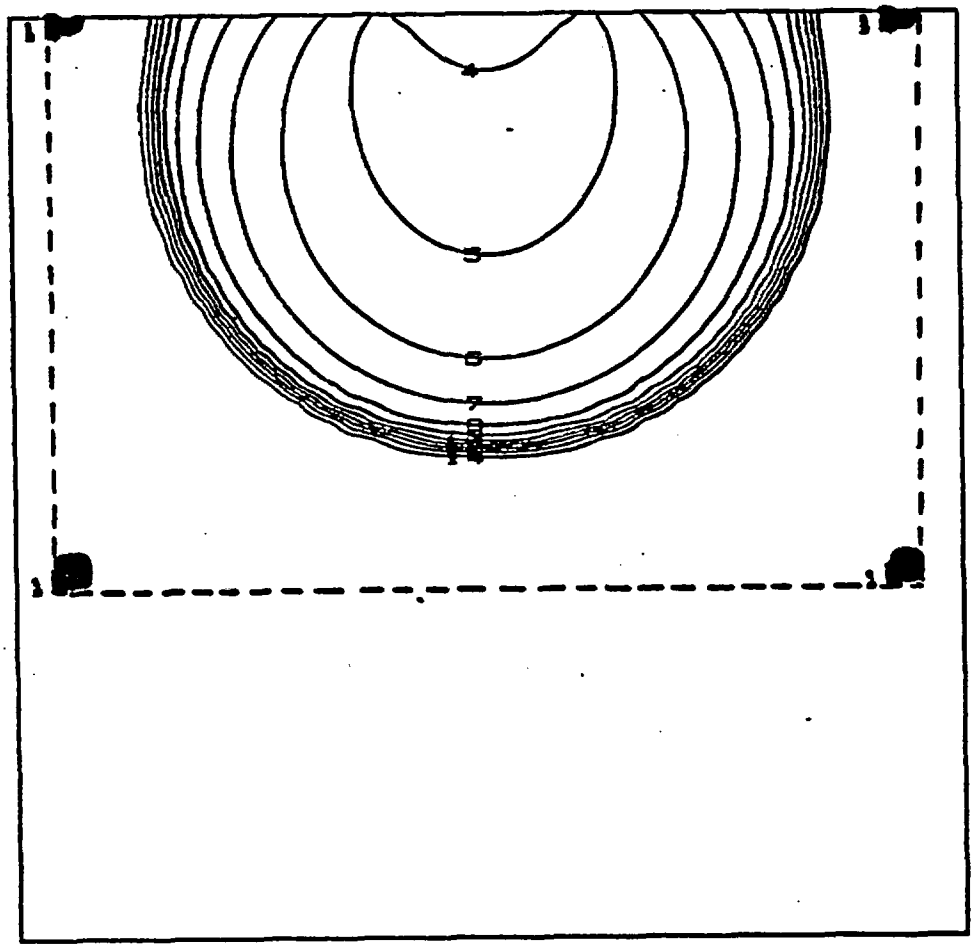


Figure 5.12 (a) Pressure head contours obtained after 1 day of infiltration with the variable domain procedure for 2D strip source infiltration q = 12 cm/day on the Dek sand with initial pressure h = - 150 cm..

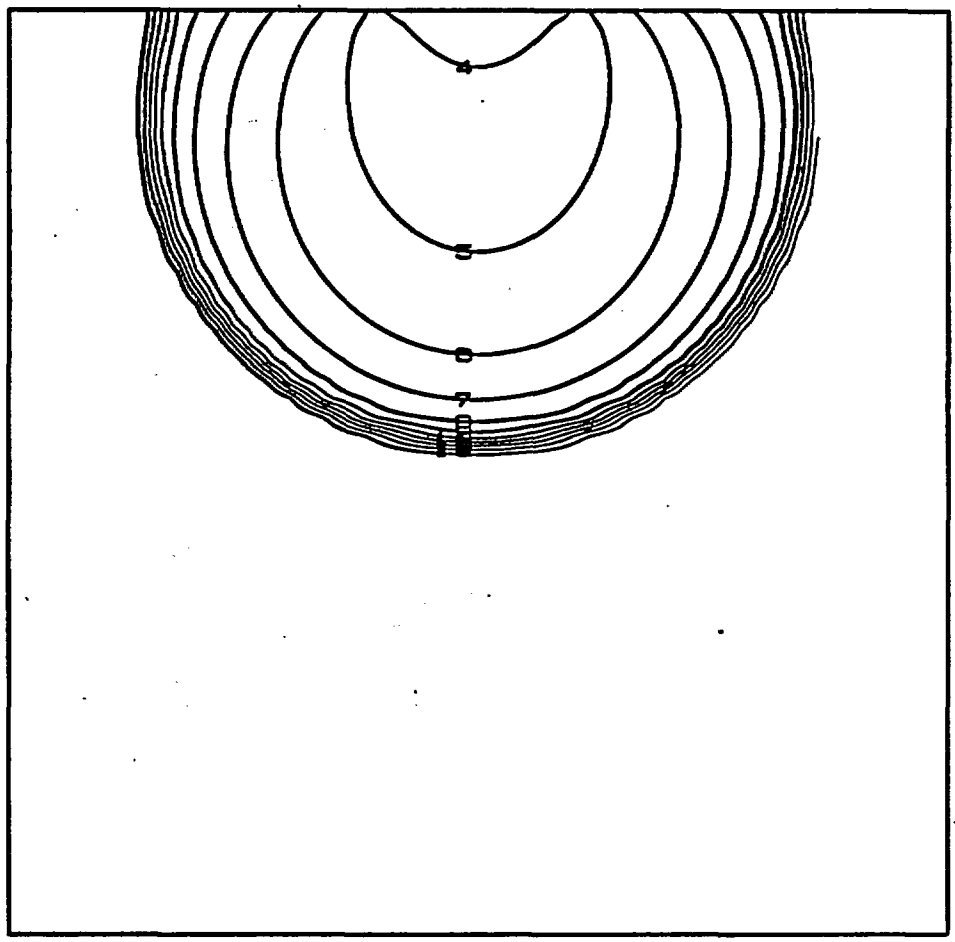


Figure 5.12(b) Pressure head contours after 1 day of infiltration with fixed domain size 150 × 150 cm and mesh size Δx = 3 cm (same case as Figure 5.12a).

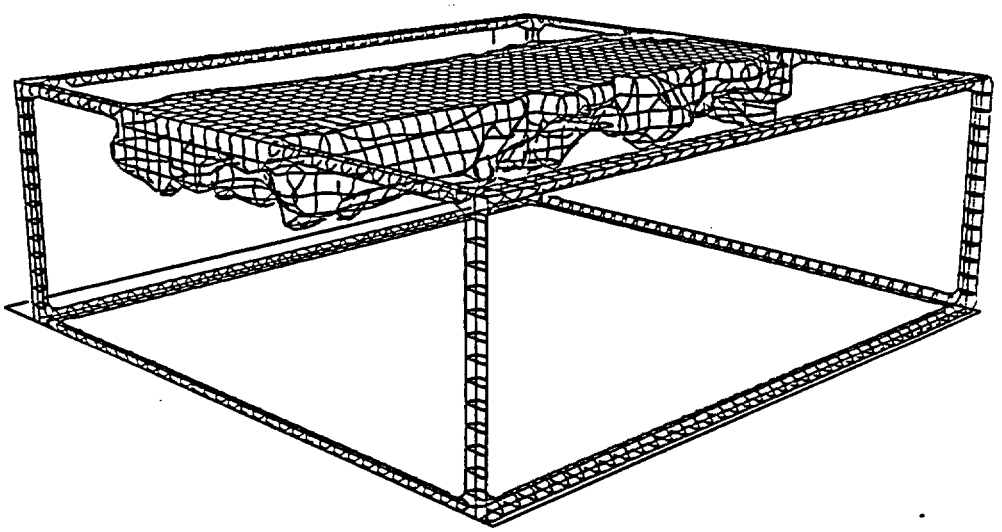


Figure 5.13 (a) Pressure head contour surface (h = - 90 cm) obtained after 1 day of infiltration with the variable domain procedure: 3D strip-source infiltration (q = 2 cm/day) on the Dek sand with random K_s and α parameter, and initial pressure h = - 150 cm

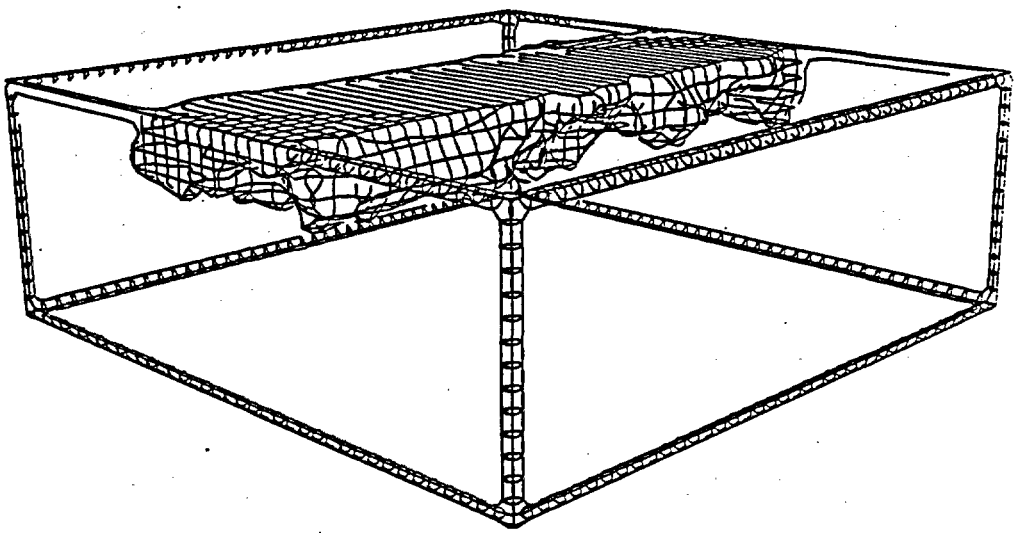


Figure 5.13 (b) Pressure head contour surface (h = - 90 cm) after 1 day of infiltration with fixed domain size 140 x 400 x 400 cm and mesh size Δx = 10 cm (same case as Figure 5.13a).

similar comparison is shown on Figures 5.13(a) and (b), this time for the case of three-dimensional strip source infiltration on a random, statistically isotropic soil whose mean properties are the same as those of Figure 5.12. The same pressure contour surface (h = -90 cm) is shown in both Figures 5.13(a) and (b): again the solutions for variable and fixed domain size seem undistinguishable visually. It should be noted however that, in the random case, the bottom boundary moved rapidly downwards to reach its prescribed maximum depth. This was due to the occurrence of non-negligible changes of pressure even far below the "wet zone".

Finally, the mass balance was computed automatically by the code at every time step based on the following algorithm. First, the computational domain for mass balance was defined as the sub-domain obtained by deleting a half-mesh size near each of the six planar boundaries (assuming here that the boundaries are fixed, for simplicity of exposition). Second, the total mass inside this subdomain was computed by integrating the volumetric moisture content $\theta(\underline{x})$ according to a simple trapezoidal rule in three dimensions. Accordingly, the mass of a node-centered cell located at node (i_1, i_2, i_3) was calculated as follows:

$$\text{Mass}(i_1, i_2, i_3) = \theta[h(i_1, i_2, i_3)] \cdot \Delta x_1 \cdot \Delta x_2 \cdot \Delta x_3 .$$

On the other hand, the total mass entering or leaving the system was obtained by summing the normal fluxes at each node of the six planar boundaries, and integrating over time. The normal fluxes were calculated according to:

$$q_{\frac{1}{2}} = - \hat{K}_{\frac{1}{2}} \cdot \left(\frac{h_1 - h_0}{\Delta x} + g\right)$$

where the index 1/2 indicates the mid-nodal location adjacent to the boundary under consideration (other indices have been omitted). In the case where all boundary conditions are of Neuman type (fixed flux), this algorithm gives the same result as would be obtained by summing directly the prescribed fluxes at the boundaries, at least within machine precision.

The accuracy of the numerical solutions was examined from the point of view of mass balance by looking at the time-dependent relative errors:

$$e(t) = \frac{Q_{mass} - (Q_{in} - Q_{out})}{Q_{in} - Q_{out}} = \frac{\delta Q}{Q} \qquad (5.134)$$

$$E(t) \simeq \frac{\int_0^t \delta Q(\tau) d\tau}{\int_0^t Q(\tau) d\tau}$$

where Q_{mass} is the rate of change of total mass in the system,

and Q_{in} and Q_{out} are the discharge rates in and out of the system (both positive quantities here). The relative error on the mass rate of change (e) usually oscillated during the early time of infiltration (this has been observed with other numerical models as well, e.g. Ababou 1981). In some of the numerical simulations presented in this work, the amplitude of oscillations could attain relatively high values (about 10%, and up to 100% in certain cases) but only for a limited number of time steps. Both error indicators e(t) and E(t) were usually very small after a sufficient number of time steps, even for fairly "difficult" cases. For instance, the relative error on total mass, E(t), was well below 1% at time t = 1 day for the infiltration problems of Figures 5.12 and 5.13, involving fixed as well as variable domain size. Figure 5.13.b in particular was for a 25,000 node grid with random soil properties.

The evolution of the relative mass balance errors e(t) and E(T) was also monitored for the more difficult infiltration problems to be analyzed in Chapter 7: see the 300,000 node grid simulations of transient strip source infiltration and steady rainfall infiltration in a random anisotropic soil (respectively sections 7.3 and 7.4). In both cases, the error on the total mass present in the flow domain rarely exceeded 10-15%. This was judged to be quite satisfactory given the high variability, nonlinearity, and large size of the system. Note that the steady

"rainfall infiltration" solution was obtained by running the flow simulator in the transient mode. The information provided by the mass balance subroutine was used to detect the convergence of the transient flow system to a steady state.

[c] Comparisons of Numerical and Analytical Solutions:

Figure 5.14 depicts the numerical solution of the 2D strip source infiltration problem (time t = 1 day) for a special type of soil having nonlinear conductivity and water retention curves, but constant "moisture diffusivity". The wet zone in that case is characterized by a fairly smooth spatial variation of pressure, and the absence of a sharply defined wetting front. Figure 5.15 compares the numerical and quasi-analytical solution obtained at a shorter time t = 0.5 day. The quasi-analytical solution was only possible because of the special form taken by the nonlinear constitutive properties of the soil, as shown below:

$$K(h) = K_s \exp(\alpha h)$$
$$\theta(h) = \theta_s \exp(\beta h) \qquad (5.135)$$
$$\alpha = \beta$$

Note that the soil moisture diffusivity corresponding to the conductivity and water retention curves (5.135) is constant,

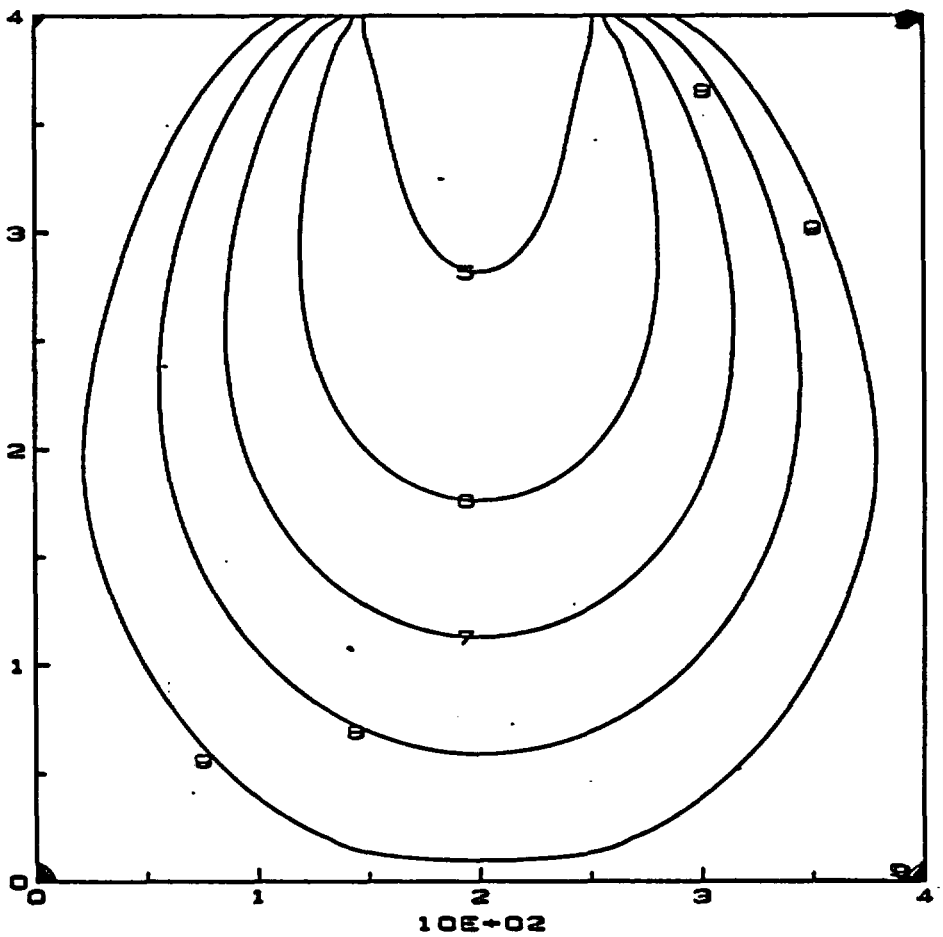


Figure 5.14 Example of numerical pressure contour map for 2D strip-source infiltration in a homogeneous soil with exponential K(h) and θ(h) curves having the same slope ($\alpha = \beta = 0.1$ cm^{-1}). Time t = 1 day.

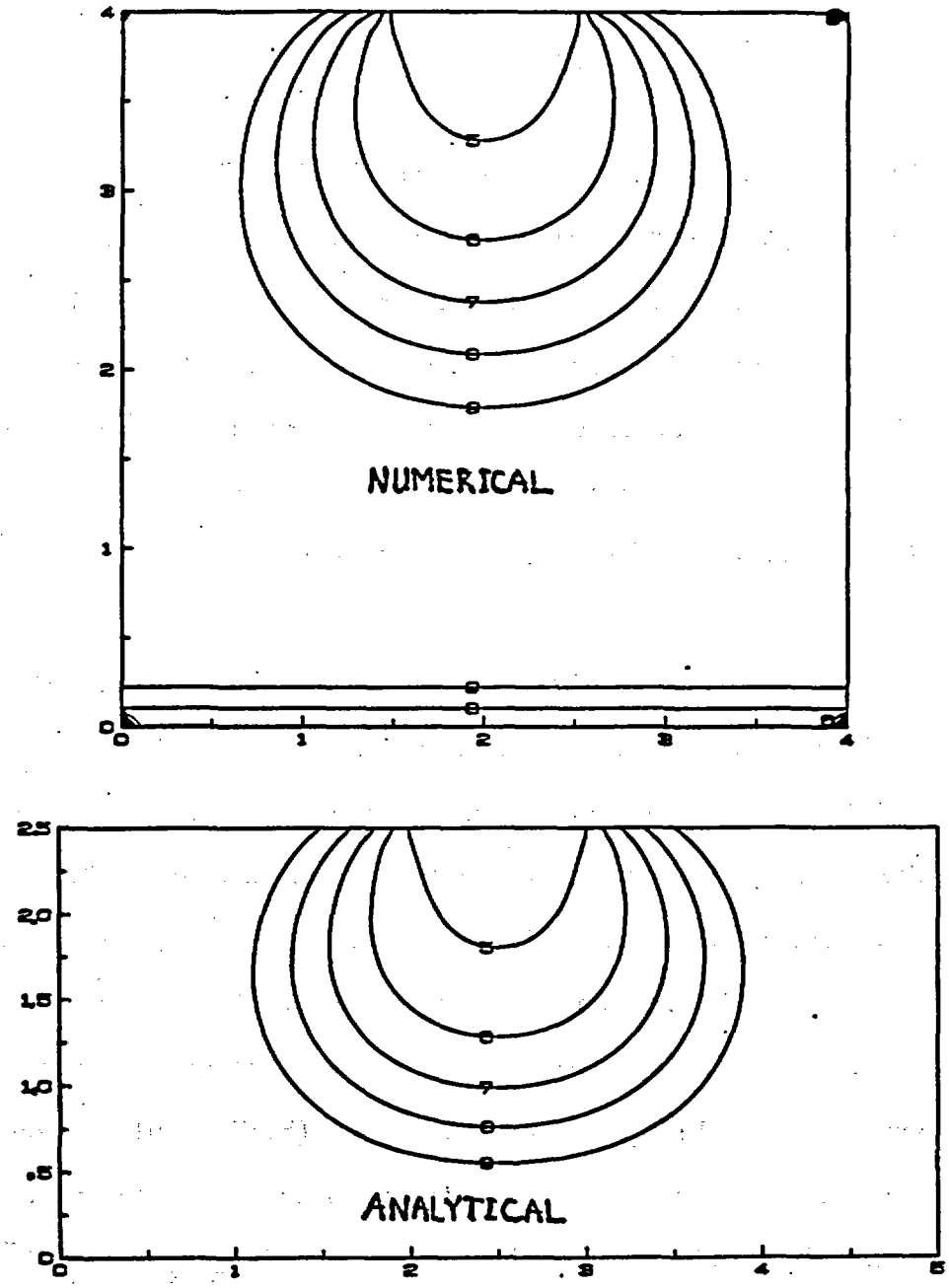


Figure 5.15 Comparison of numerical and analytical solutions for 2D strip-source infiltration in a homogeneous soil with exponential K(h) and θ(h) curves having the same slope ($\alpha = \beta = 0.1$ cm^{-1}). Pressure contours at time $t = 0.5$ day

independent of pressure:

$$D = K_s/(\alpha\theta_s).$$

As a consequence, the unsaturated flow equation expressed in terms of the conductivity or Kirchhoff transform (Equation 5.131) becomes linear. Warrick and Lomen (1976) developed quasi-analytical solutions for the case of strip-source and disc-source infiltration under constant flux. The particular program to calculate the strip-source solution shown in Figure 5.15 was developed by us (Ababou, 1981).

Unfortunately, the soil properties (5.135) are not realistic enough to obtain a reasonable simulation of transient infiltration phenomena, due to the fact that in general the diffusivity is far from constant (see Ababou, 1981). Another drawback is that the solution given by Warrick and Lomen (1976) is exact only in the limit of zero initial conductivity. The slight discrepancy that can be observed between the numerical and analytical solutions shown in Figure (5.15) could be due to the fact that the initial conductivity in this example was not really negligible relative to the input flux ($K_{in}/q_o = 4.5\ 10^{-3}$). In addition, the quasi-analytic solution requires a numerical evaluation of integrals of special functions; the computed

pressure contours close to the initial state (far from the source) are quite sensitive to small errors of numerical integration.

Nevertheless, the agreement between the numerical and analytical solutions of Figure (5.15) seems reasonably good, especially close to the source. The two-dimensional numerical solution shown on top was obtained by shrinking the longitudinal domain size to just 5 nodes (3 internal nodes) and picking the central slice for visual display. The near-perfect symmetry of the numerical solution is an indication that the nonlinear-SIP solver worked well in that case: it should be noted indeed that the three-dimensional SIP solver is inherently assymetric. This asymmetry would probably show up in cases of incomplete convergence.

Figure 5.16 shows the result of another comparison between numerical and analytical solutions. In this case, the flow simulator was used in both the "saturated" and "unsaturated" mode to solve the one-dimensional linear diffusion equation:

$$\frac{\partial H}{\partial t} = D \cdot \frac{\partial^2 H}{\partial x^2} \quad , \quad 0 \leq x \leq L$$

$$H(0,t) = 1 \qquad (5.136)$$

$$H(L,t) = 0$$

$$H(x,0) = 0.$$

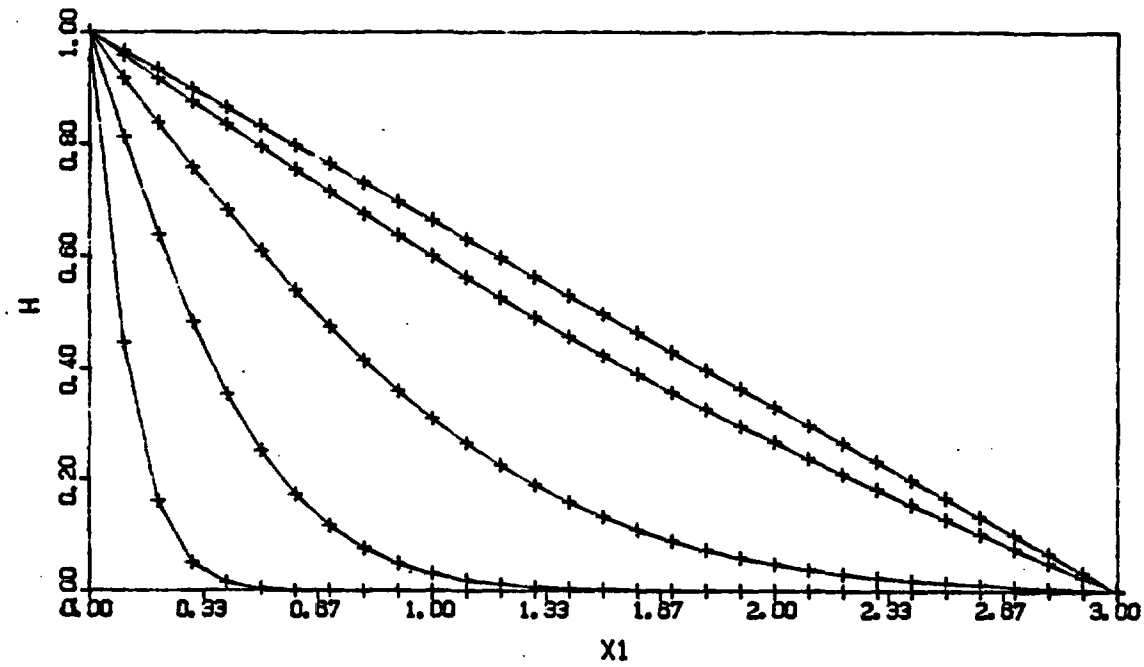


Figure 5.16 Numerical and analytical solutions for the transient 1D diffusion equation with constant coefficients. The numerical solutions obtained in the saturated or unsaturated modes, with fixed or variable time steps, were undistinguishable from the analytical solution. One of the numerical solutions is shown here for times $t = 0.01$, 0.10, 0.5, 1, and $t \geq 5$ (quasi-steady state)

This equation models for instance the transient recharge of a confined aquifer with conductivity K and specific storativity S, such that D = K/S.

The analytical solution of equation (5.136) can be expressed in the form of an infinite series with sine and exponential functions (Korn and Korn, 1968, p. 325). However we have found that a very high machine precision would be needed to obtain reasonably accurate answers at early times. Another series solution was finally worked out by using a superposition of Green's functions as explained by Godunov (1973, pp. 29–41). The final result is given below in dimensionless space-time variables:

$$H(y,\tau) = \sum_{n=1}^{\infty} n\cdot(a_n - b_n)$$

$$a_n = \operatorname{erf}\left(\frac{n+y/2}{\sqrt{\tau}}\right) + \operatorname{erf}\left(\frac{n-y/2}{\sqrt{\tau}}\right) \qquad (5.137)$$

$$b_n = \operatorname{erf}\left(\frac{n-1+y/2}{\sqrt{\tau}}\right) + \operatorname{erf}\left(\frac{n+1-y/2}{\sqrt{\tau}}\right)$$

where:

$$y = x/L$$

$$\tau = D\cdot t/L^2$$

and erf(x) is the usual error function defined by:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \cdot \int_0^x e^{-s^2}\cdot ds.$$

The numerical solutions obtained with both the saturated and unsaturated options of the flow simulator for a three-dimensional elongated domain, fitted perfectly the one-dimensional analytical solution computed from equation (5.137). Figure (5.16) showed only one graph of H(x,t) because the three solutions were visually indistinguishable. It should be noted that the saturated flow mode was implemented with constant time steps, while the unsaturated mode was implemented with a variable time step size. In the latter case, the constant coefficients were obtained by taking K(h) constant and θ(h) linear. This test was useful to check the soundness of various algorithms of the flow simulator, including the variable time step procedure and the performance of the SIP matrix solver for transient problems. In other words, this particular test problem provided an accurate check on many features of the flow simulator, other than the nonlinear solver.

[c] Infiltration Experiments with Homogeneous and Layered Soils:

We now proceed to analyze, in a rather qualitative way, the numerical pressure fields obtained for a few test problems of one and two-dimensional infiltration in uniform or deterministically layered soil systems. Our intent here is merely to demonstrate that, in all cases considered, the numerical solution agrees with intuitive and/or physically based

principles. We consider first a few cases of infiltration in uniform soils, and continue our exploration with the case strip-source infiltration in two-dimensional deterministically layered soils. In the process, one of these test problems will be used to examine the effect of mesh size on the numerical solution.

The first sequence of test problems concerned one and two-dimensional infiltration on the homogeneous "Dek" sand with a relatively dry initial state (h_{in} = -111 cm). The constitutive relations of this soil were given earlier in Figure (5.3). The pressure profiles obtained for one-dimensional infiltration in a 3 meter deep soil column, with zero pressure at the top, are depicted in Figure (5.17). The resolution Δx = 3 cm was fine enough to capture the very sharp wetting fronts obtained in this case. The next two Figures (5.18) and (5.19) display the vertical pressure profile and the pressure contours obtained for 2D infiltration with zero pressure maintained at the surface of a strip-source (saturated strip). The wetting front is still very sharp, although less so than for the one-dimensional problem. In addition, it can be seen that that front moves downwards at a lesser rate due to increased dimensionality (lateral diffusion). Finally, Figures (5.20) and (5.21) display the vertical pressure profiles and the pressure contours obtained for 2D strip-source infiltration with a fixed flux (q_o = 12 cm/day) at the surface of

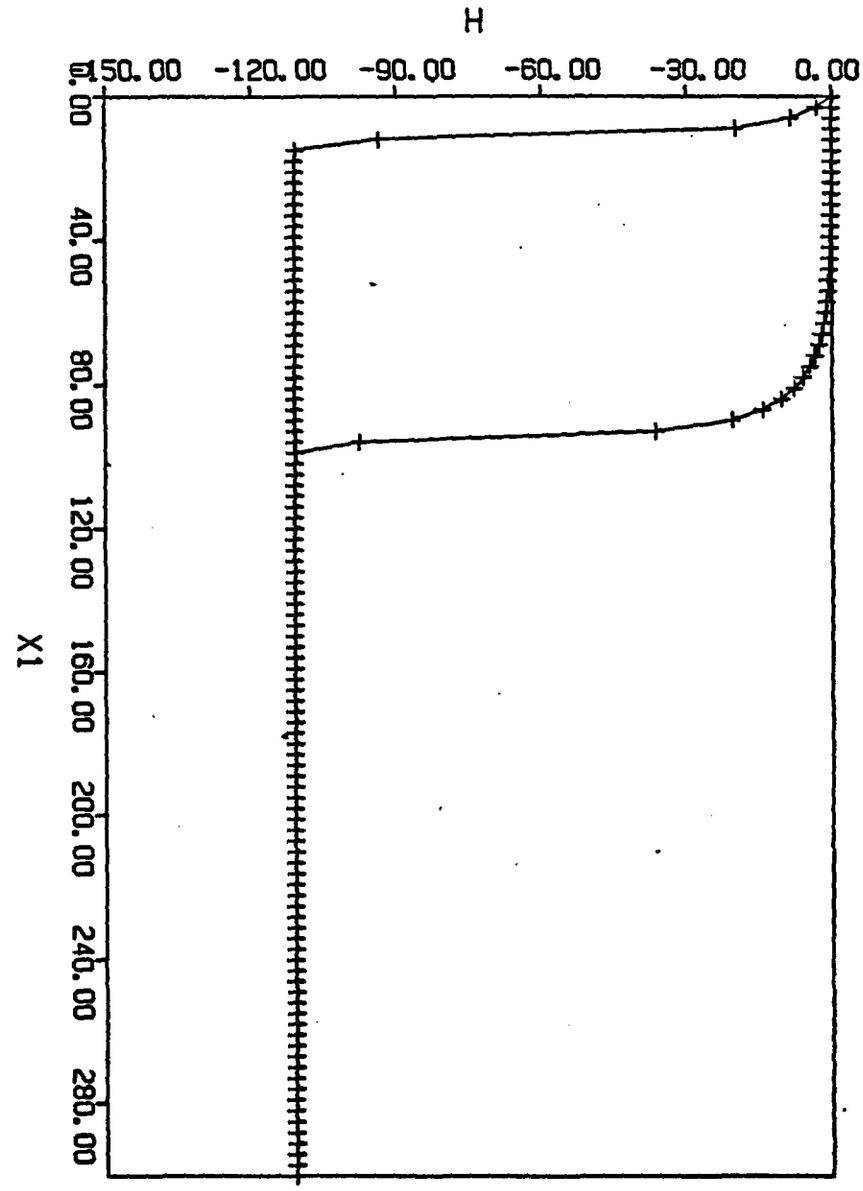


Figure 5.17 Vertical pressure profile at times t = 0.005 and 0.1 day for one-dimensional infiltration with zero pressure at soil surface (Dek sand with h_{in} = - 111 cm). The vertical mesh size is Δx = 3 cm and the total length of the column is L = 300 cm.

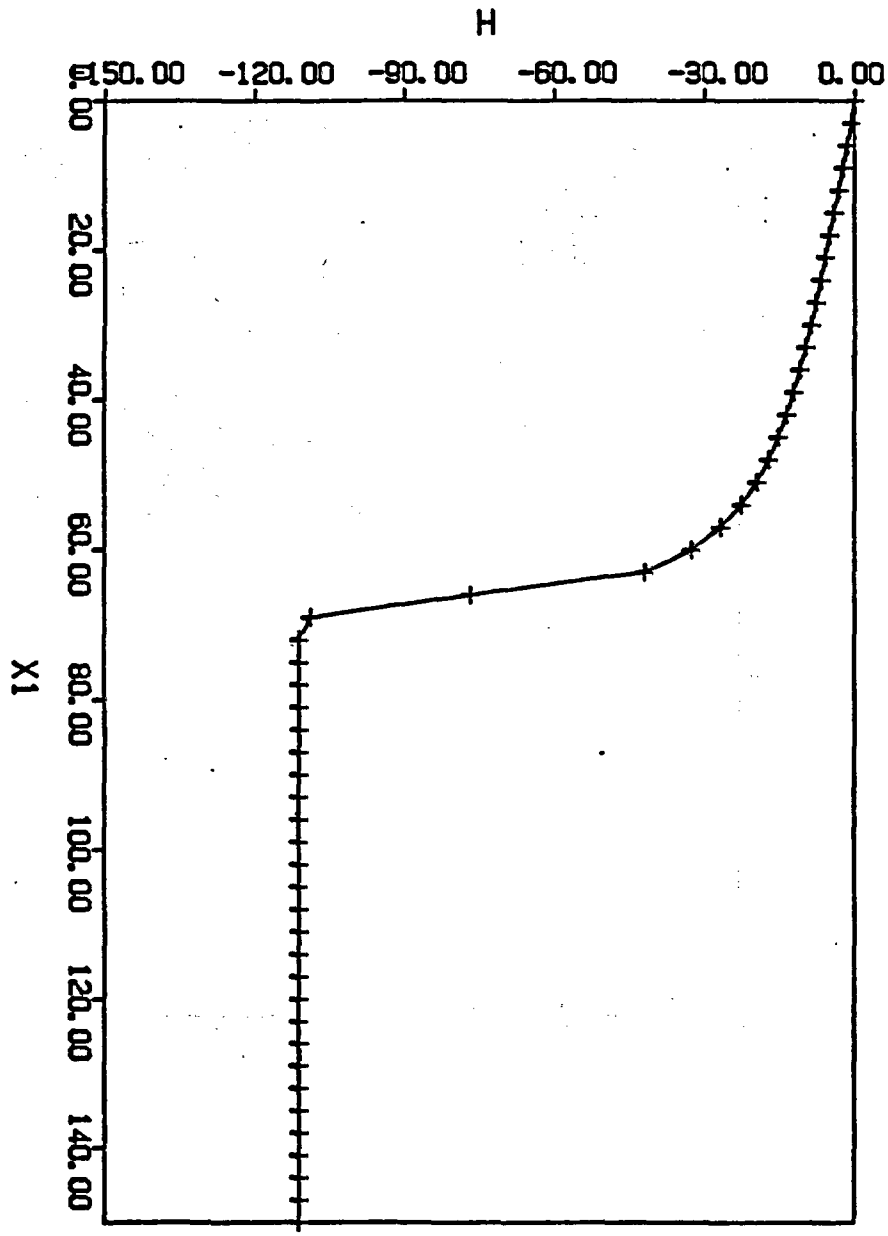


Figure 5.18 Vertical pressure profile at time t = 0.1 day for a two-dimensional infiltration with a saturated strip source. The vertical transect coincides with the axis of symmetry (see Figure 5.19).

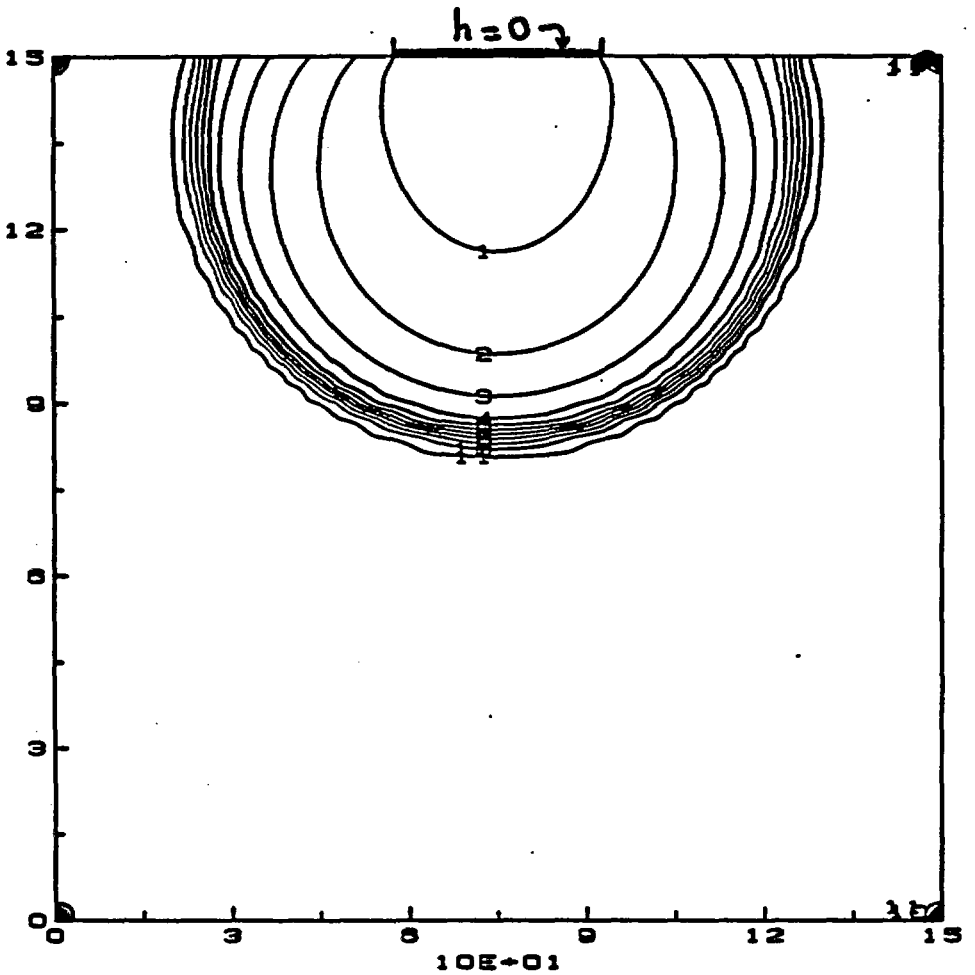


Figure 5.19 Pressure head contour lines at time t = 0.1 day for 2D infiltration with a saturated strip source (Dek sand with h_{in} = - 111 cm). The source width is 33 cm, the mesh size is 3 cm, and the domain size 150 x 150 cm.

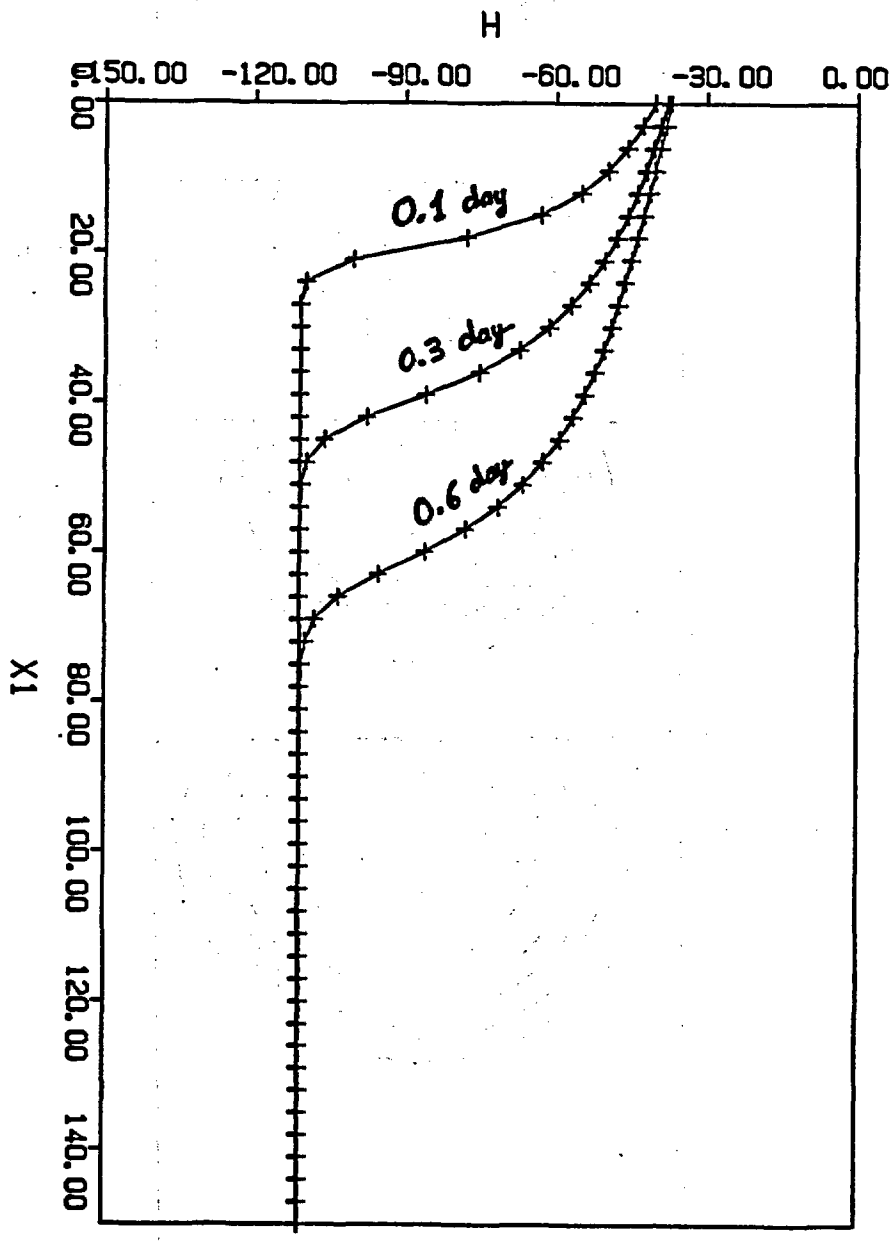


Figure 5.20 Vertical pressure profiles at times t = 0.1, 0.3 and 0.6 day for a two-dimensional infiltration with a constant flux strip-source. The vertical transect coincides with the axis of symmetry (see Figure 5.21)

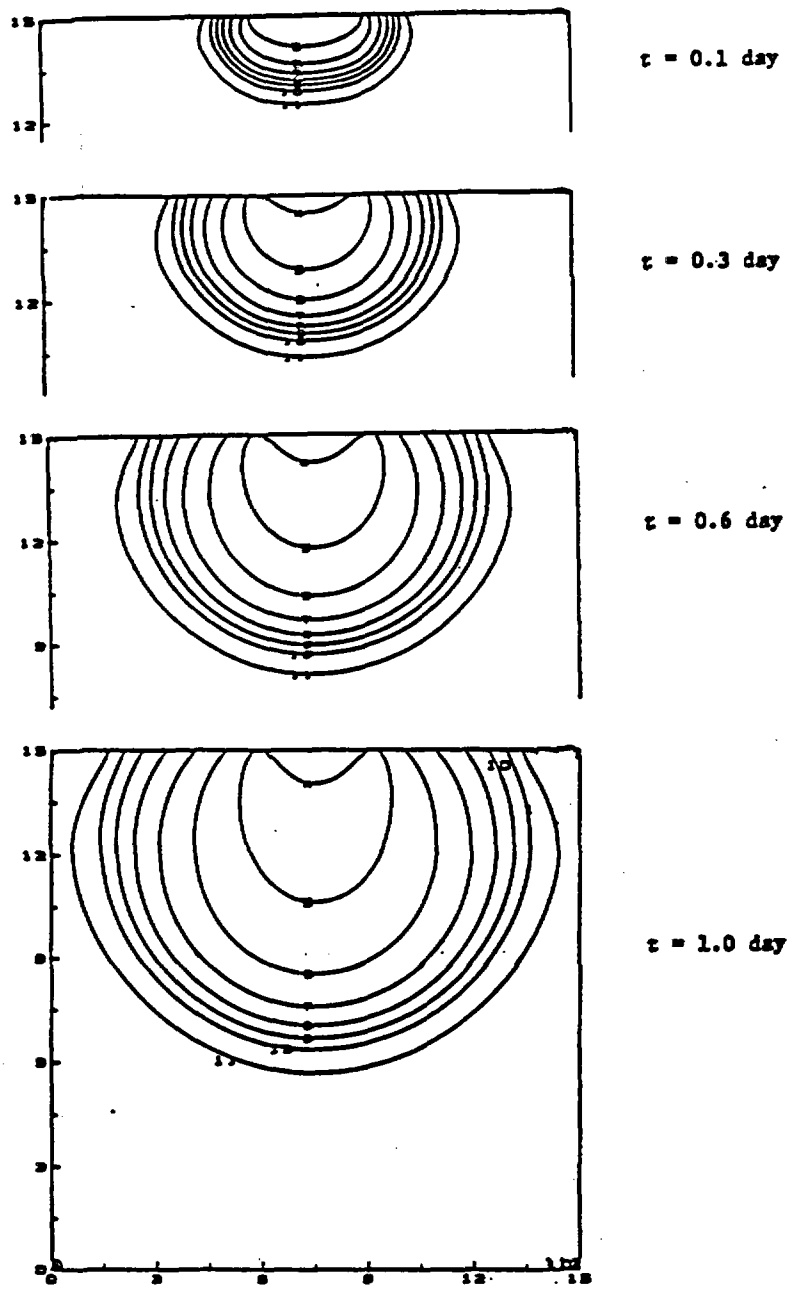


Figure 5.21 Pressure head contour lines for 2D infiltration with a constant flux strip source q_o = 12 cm/day (Dek sand with h_{in} = - 111 cm). The source width is 33 cm, the mesh size 3 cm, and the domain size 150 x 150 cm. Times: t = 0.1, 0.3, 0.6 and 1.0 day

the strip. In this case, the wetting front appears even smoother, and the rate of growth of the wet zone is slower than in the previous case of a saturated strip (in both cases, the strip width was 33 cm). Note that the prescribed flux was 8.5% the value of the saturated conductivity, whereas the flux over the saturated strip source in the previous example was presumably much larger. Last but not least, we emphasize the fact that the numerical solutions obtained for the strip-source problems seem perfectly symmetric about the central vertical axis. This feature was not built-in the solution, but rather resulted from the convergence of the nonlinear-SIP solver towards the exact solution; it should be kept in mind indeed that the SIP factorization is not symmetric, so that the numerical solution would probably appear non-symmetric in case of incomplete convergence.

In order to explore the problem solving capabilities of the flow simulator for spatially variable unsaturated soils, we have simulated two-dimensional strip-source infiltration on horizontally and vertically layered soils. This could be viewed as an intermediate case between the ideal case of homogeneous soils, and the more realistic case of three-dimensional *random* soils to be analyzed in Chapter 7 (statistically layered soils in particular). Of course, we expect that the flow patterns will be

much easier to analyze for uniformly layered soils than for statistically layered (random) three-dimensional soils. That is precisely the reason for our choice here.

Three types of uniformly layered soil systems were considered for qualitative analysis: a horizontally layered system with mild contrast (sand/sand); another horizontally layered system with high contrast (sand/silt); and a vertically layered system with high contrast (sand/silt). Briefly, the sand/sand system corresponds to alternate layers of the Dek sand of Figure 5.3, and a somewhat coarser sand (Dieri sand, Ababou 1981). The sand/silt system corresponds to alternate layers of the Dek sand of Figure 5.3 and the Montfavet silt of Figure 5.4. The contrast for each layered system can be characterized by an index of variability of K_s and α, the two parameters of the exponential K(h) curve, as follows:

$$\sigma_{\ell n\ Y} = \left\{ \frac{(\ell n\ Y_1/Y_G)^2 + (\ell n Y_2/Y_G)^2}{2} \right\}^{1/2}$$

$$Y_G = \sqrt{Y_1 \cdot Y_2}$$

where Y is either K_s or α, and the index (1,2) refers to the two soils composing the layered system. For the sand/sand system, the contrast was moderate:

$$\sigma_{\ell n\ K_S} \simeq 0.44$$

$$\sigma_{\ell n\ \alpha} \simeq 0.22$$

and, for the sand/silt system the contrast was quite high:

$$\sigma_{\ell n\ K_s} \simeq 2.60$$

$$\sigma_{\ell n\ \alpha} \simeq 0.87$$

Figure (5.22) shows pressure head contour lines obtained at time t = 1 day for the sand/sand system with alternate horizontal layers of thickness 9 cm. The initial pressure was h = -150 cm for figure (a), and h = -90 cm for figure (b). The general shape of the wetted zone, apart from small scale oscillations, is quite similar to that obtained for either sand soil alone, that is, without alternate layering. But, for the highly contrasted sand/silt system shown in Figure (5.23), the pressure field looks quite different. In this case, the initial pressure (h = - 150 cm) was such that the initial conductivity was higher in the silt than in the sand, by two orders of magnitude. Consequently, as we have observed from detailed numerical outputs, the wetting front rested most of the time just above a sand layer, while moisture spread laterally within the silt layers.

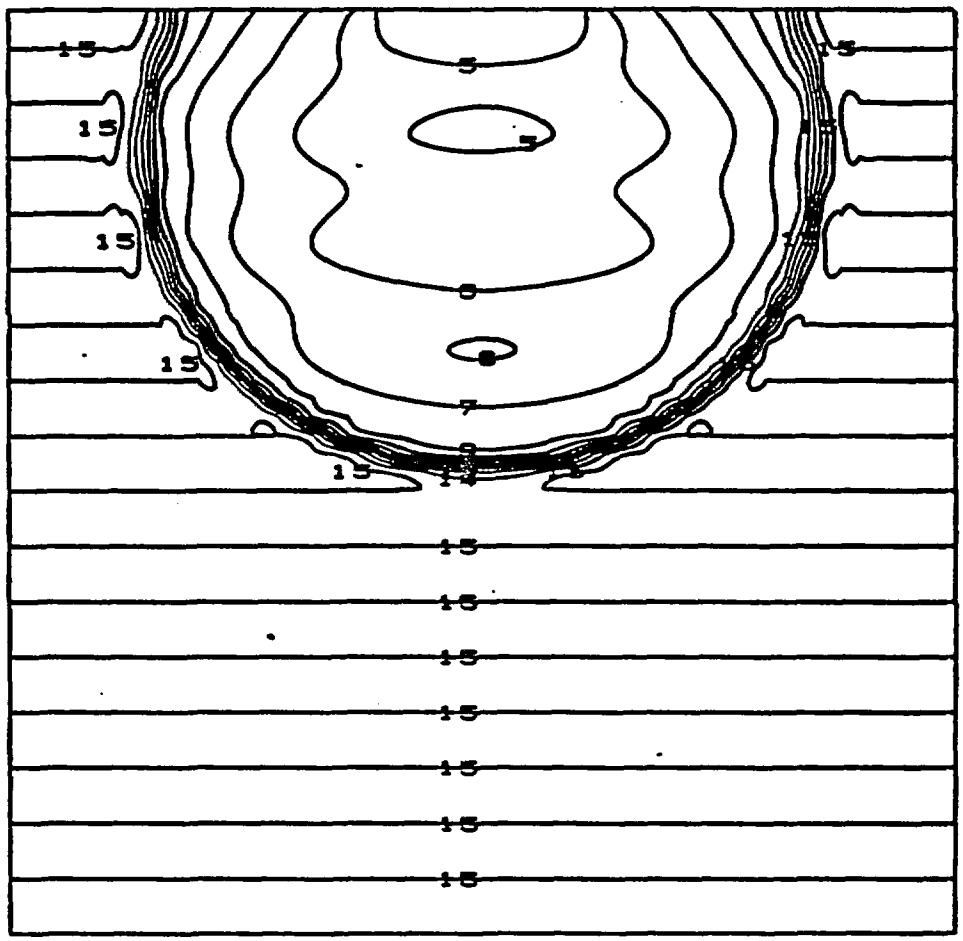


Figure 5.22 (a) Pressure head contour lines for 2D strip-source infiltration (q = 12 cm/day) in a horizontally layered sand/sand system, at time t = 1 day. The mesh size is 3 cm, the domain size 150 x 150 cm, the strip width 33 cm, and the alternate layers thickness 9 cm. The initial pressure head was h_{in} = - 150 cm

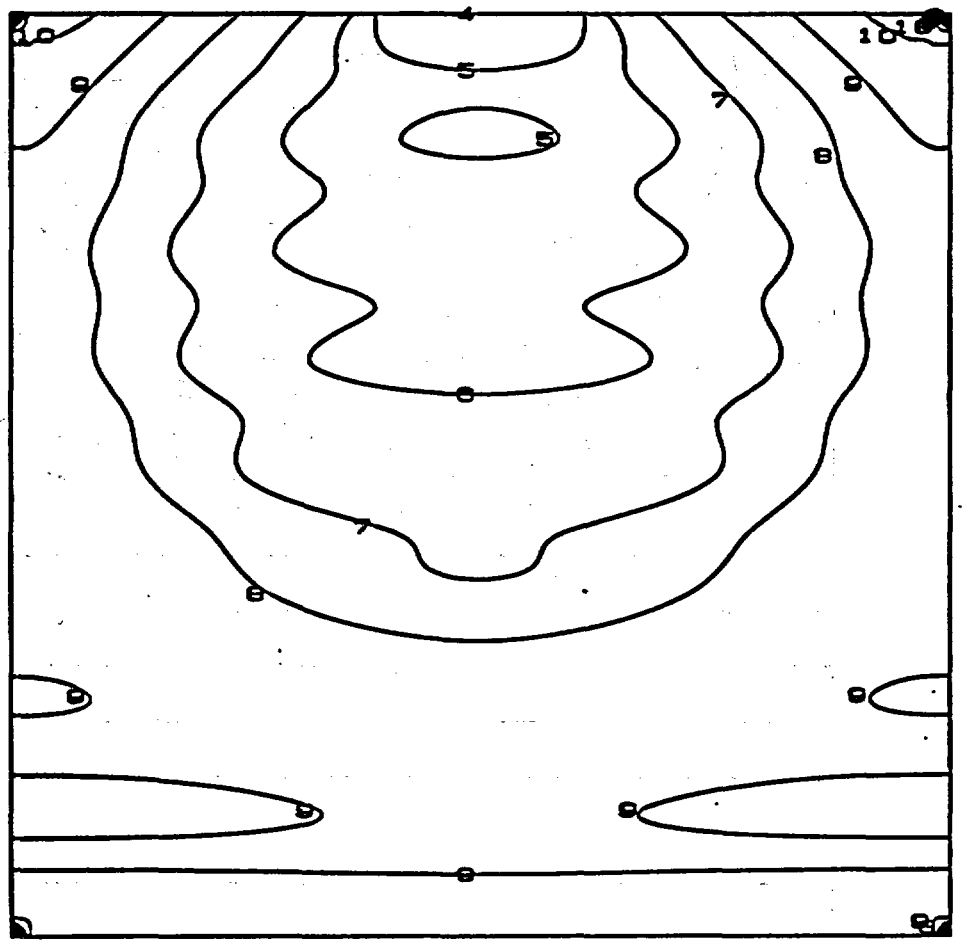


Figure 5.22(b) Same as Figure 5.22.a, but with a less dry initial state (h_{in} = - 90 cm).

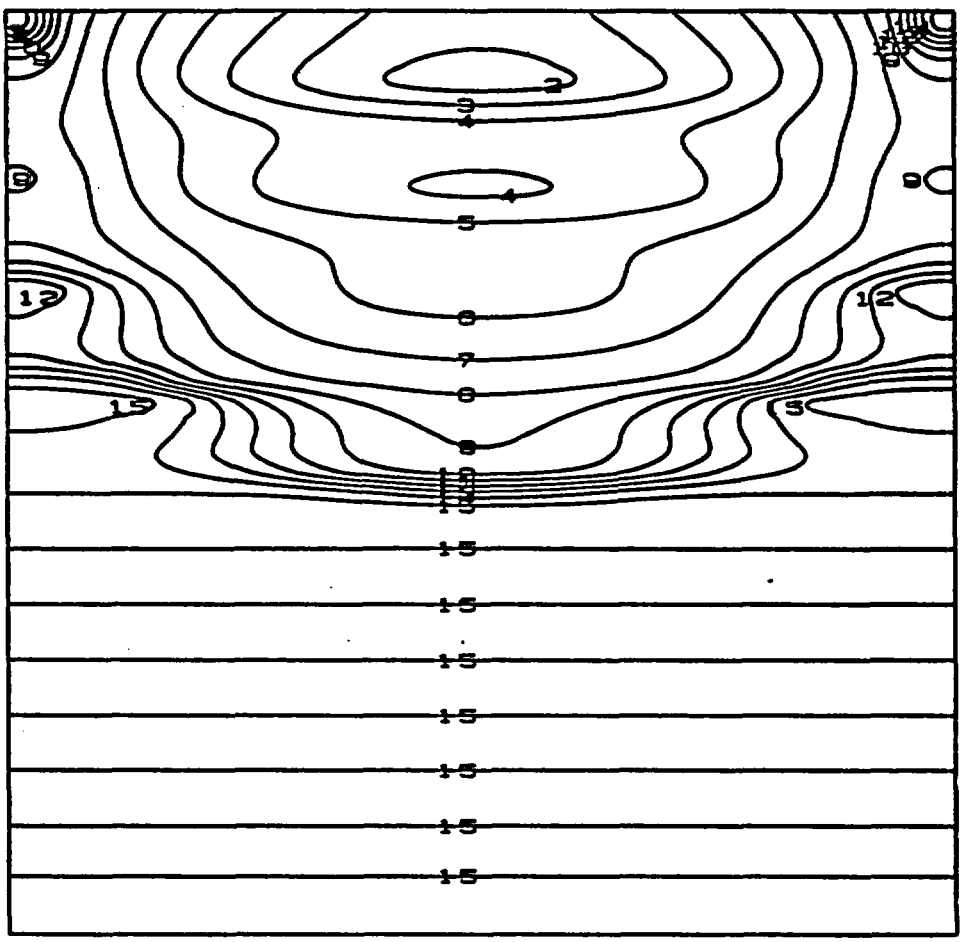


Figure 5.23 Pressure head contour lines for 2D strip-source infiltration (q = 12 cm/day) in a horizontally layered sand/silt system, at time t = 1 day. The mesh size is 3 cm, the domain size 150 × 150 cm, the strip width 33 cm, and the alternate layers thickness 9 cm. The initial pressure was h_{in} = - 150 cm.

It should be noted that the mesh size in these examples was one third of the layers thickness, i.e., $\Delta x = 3$ cm. Figure (5.24) shows the solution obtained for the sand/silt system with a coarser grid equal to the layer thickness, i.e., $\Delta x = 9$ cm. The obvious effect of the coarse mesh is that it smears out the small scale fluctuations of pressure obtained with the finer grid (compare Figures 5.23 and 5.24). However, the overall shape and size of the wet zone are surprisingly well represented with the coarse grid simulation. This can be seen more easily by representing on the same plot the vertical pressure profiles obtained with the fine and coarse meshes (Figure 5.25). This relative agreement indicates that the grid resolution need not be much finer than the typical layer thickness in order to obtain realistic solutions. The grid Peclet number constraint (5.128) should be also kept in mind. In the present case, the coarse grid Peclet number (with $\Delta x = 9$ cm) was about 0.6 for the sand layers, and 0.1 for the silt layers. Both these values satisfy the "nonlinear" stability constraint $Pe \lesssim 2$.

Finally, Figure (5.26) shows the pressure head contour lines obtained for the *vertically layered* sand/silt system, with alternate layers of thickness 9 cm. The non-symmetrical shape of the pressure field is due to the fact that the vertical axis

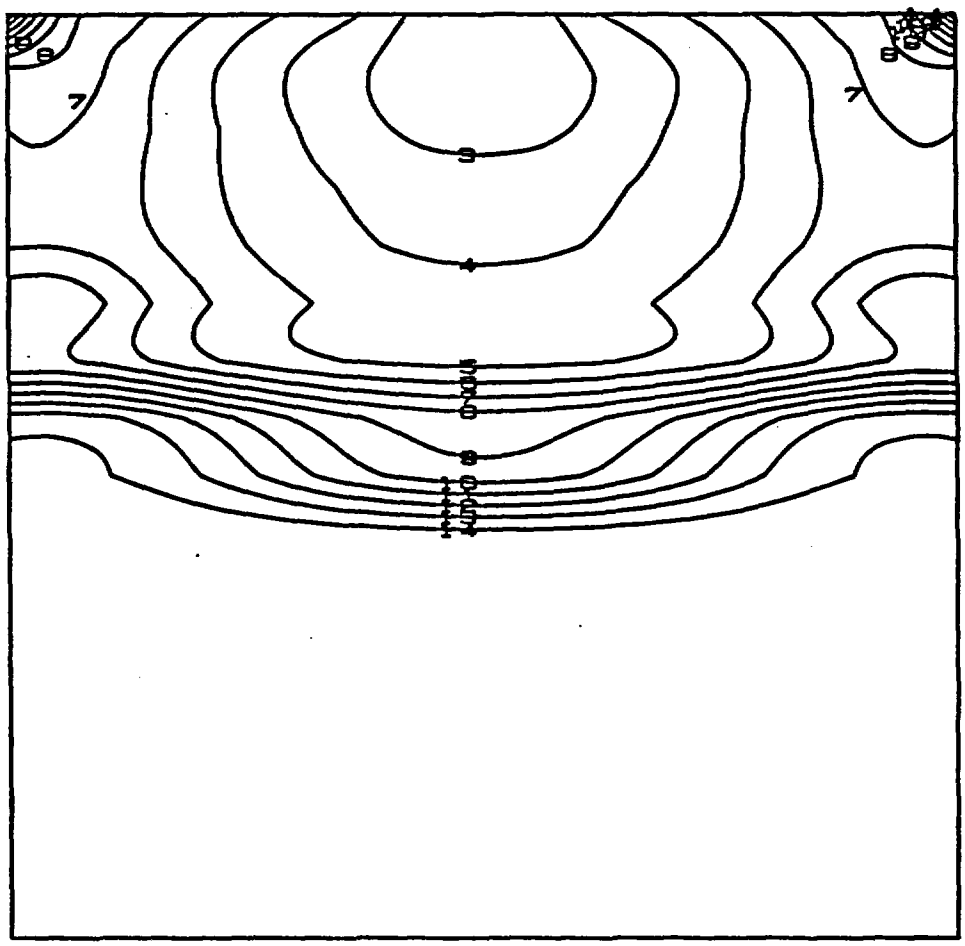


Figure 5.24 Same as Figure 5.23, but with a coarser mesh size Δx = 9 cm equal to the layer thickness.

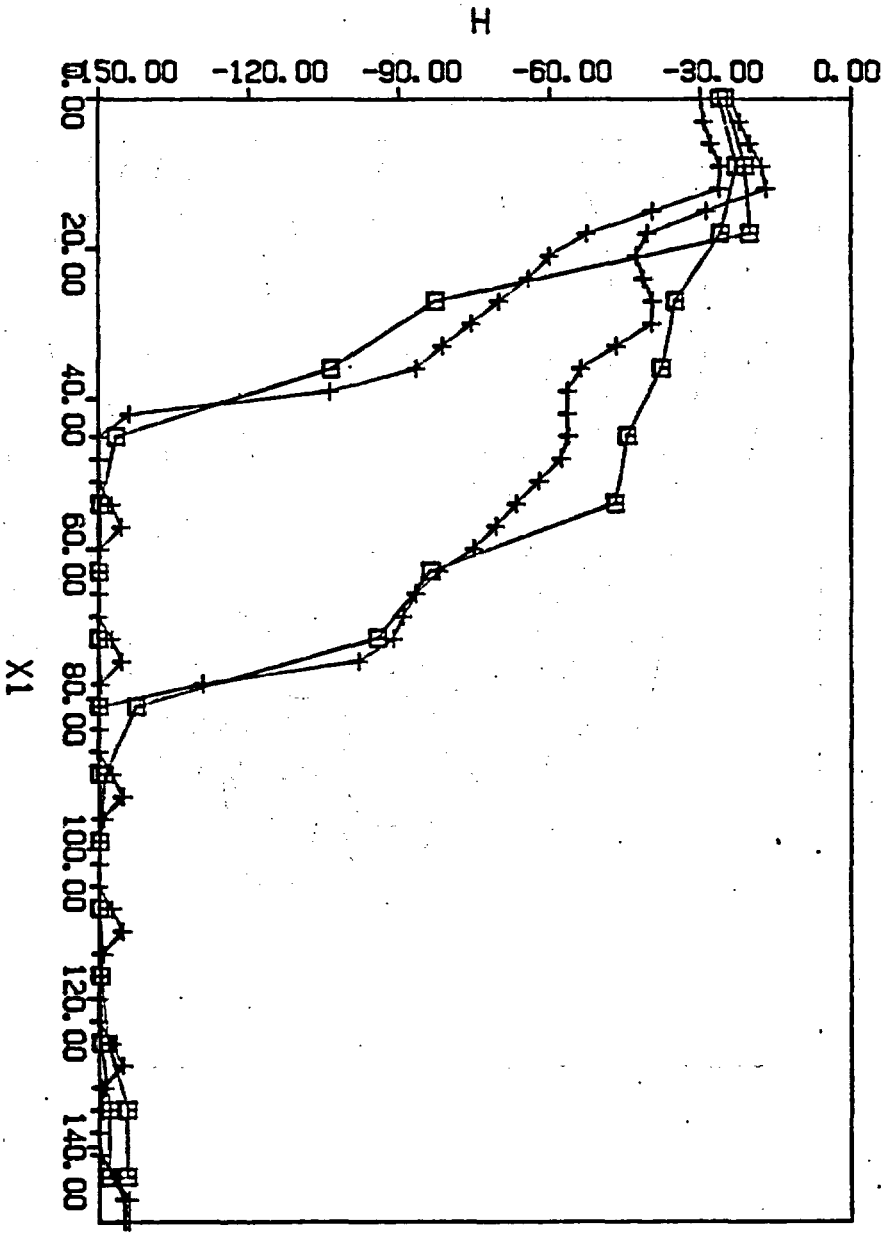


Figure 5.25 Vertical pressure profiles through the axis of symmetry of the strip source for the sand/silt system at times t = 0.3 and 1 day. The crosses correspond to the fine mesh simulation (Figure 5.23 with Δx = 3cm) and the square boxes to the coarse mesh simulation (Figure 5.24 with Δx = 9 cm).

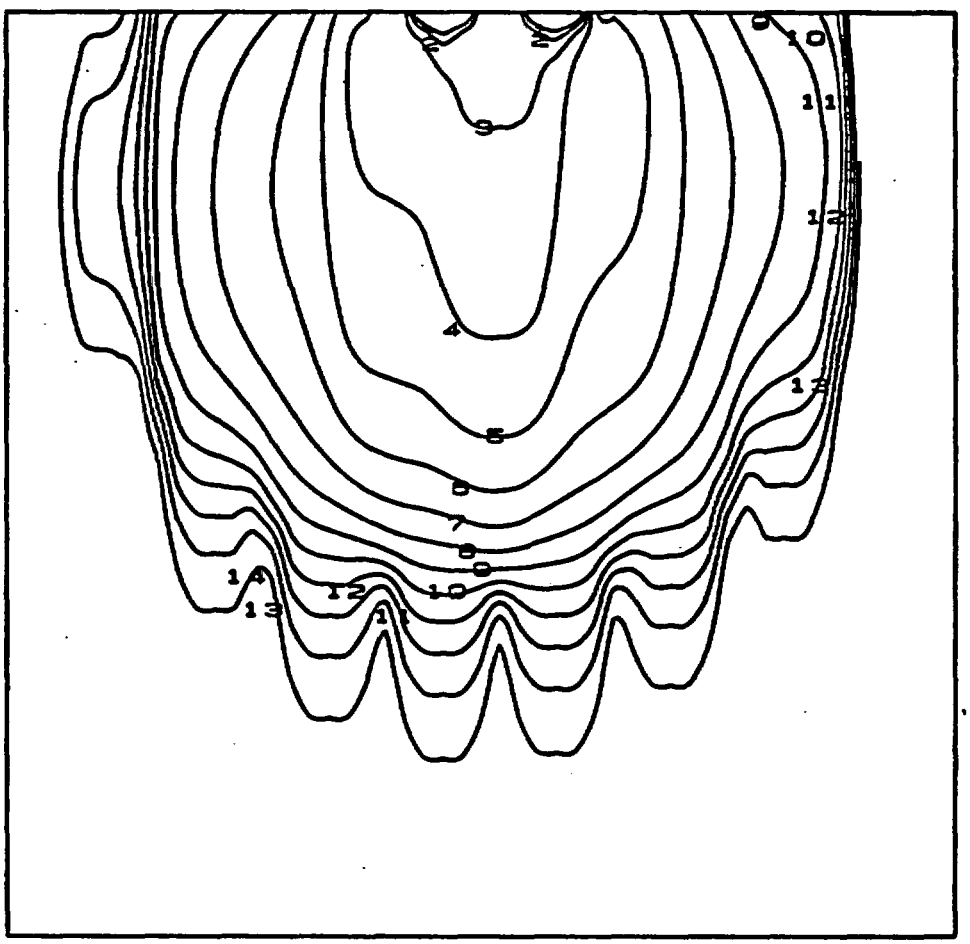


Figure 5.26 Vertically layered sand/silt soil system (strip source infiltration-time = 1day).

located at the mid-point of the strip-source does not constitute an axis of symmetry with respect to the vertical layers. One interesting feature in Figure (5.26) is that the spatial fluctuations of the wet pressure contours are more or less in opposite phase with the dry pressure contours (looking through vertical lines). This is due to the existence of a cross-over point (h = -90 cm) below which $K_{silt} < K_{sand}$, and above which $K_{silt} > K_{sand}$. Thus, it should not be surprising to see that the particular contour line h = -90 cm corresponding to $K_{silt} = K_{sand}$ is almost perfectly smooth !

5.5 Summary and Conclusions on Numerics:

In this chapter, we have developed a multi-faceted analysis of the numerical issues related to the direct simulation of large and complex flow systems. It may be useful to briefly summarize the various approaches that were developed and the conclusions that were drawn from these analyses. To begin with, we developed in Section 5.1 the basic equations for the *finite difference* approximation of both saturated and unsaturated flow phenomena in spatially heterogeneous media. The particular choice of the finite difference model was motivated by considering the likely numerical requirements for simulating representative single realizations of random flow systems (high

resolution and large domain).

These considerations were confirmed by a more systematic theoretical analysis of the stochastic finite difference solution error arising from truncation errors (Section 5.2). The remarkable result obtained for the case of steady saturated flow, i.e., for the stochastic "heat equation" with 3D random field conductivities, was that the finite difference scheme is indeed a *consistent approximation* of the stochastic equation in the mean-square sense. Furthermore, it was shown that the *order of accuracy* is $O(\Delta x/\lambda)^2$ for the hydraulic head and $O(\Delta x/\lambda)$ for the flux vector, in the case of a *smooth* log-conductivity field with correlation length λ (e.g., random field with a Gauss-shaped spectrum). On the other hand, we found that the order of accuracy drops to $O(\Delta x/\lambda)^{3/2}$ and $O(\Delta x/\lambda)^{1/2}$ for the head and flux, in the case of a *noisy* log-conductivity such as the 3D Markov field with exponential covariance function. It is remarkable that the finite difference approximation is still consistent in this case, despite the fact that the log-conductivity is non-differentiable in the mean-square sense. These encouraging results were refined further by evaluating explicitly the leading order terms of the root-mean-square errors in the head and flux. It was found that both errors were proportional to the standard deviation of the log-conductivity field. One major conclusion from this

statistical truncation error analysis is that a relatively fine grid resolution is needed to obtain even moderately accurate solutions in terms of the flux vector field. Thus, when the log-conductivity is the "noisy" three-dimensional Markov field, the root-mean-square error on the flux is as large as 15-20% for $\Delta x/\lambda = 1/3$, and still about 10% for $\Delta x/\lambda = 1/10$.

However, it was also recognized that the numerical errors of the random flow simulator will be due in part to the difficulty of solving accurately very large matrix systems. Section 5.3 was devoted to the development of an adequate linear system solver for large sparse matrices. Our literature review focused on iterative solvers, and particularly the SIP and ICCG solvers based on approximate factorizations. The SIP solver was finally chosen for implementation and was described in some detail. The accuracy of the linear system solutions obtained with SIP were analyzed in a semi-empirical way by examining the rate of convergence of the SIP iterations both from a theoretical and "experimental" point of view.

Our conclusions, based on numerical experiments for large random systems of saturated flow (up to 1 million nodes), were quite favourable. It was found that the root-mean-square solution error could be reduced to very small values (typically less than 1%) in a few hundred up to one thousand iterations,

depending on the variability of the input log-conductivity field. The numerical simulations were carried out on a Cray 2 supercomputer, requiring CPU times of one to several hours for the most "difficult" random flow problems (1 million nodes). A Microvax 2 machine was used for medium size problems on the order of 1 to 2 hundred thousand nodes. One remarkable aspect of the proposed method of analysis was that the "*true*" solution error was evaluated indirectly by using information from the actual simulations (sequence of residual errors, and convergence rate). It was shown that in many cases the true error could be much larger than the apparent (residual) error, particularly for large and noisy systems where underrelaxation was needed to achieve convergence. Empirical analysis suggested that the number of iterations required to solve linear random flow problems could be proportional to n, the largest unidirectional size of the three-dimensional rectangular grid. However, it is still not clear at this time whether the SIP solver or any similar iterative solver will actually converge for very large random flow systems on the order of 10 million equations or more. Extrapolation of our results suggested that, in case of convergence, the solution of a highly variable saturated flow problem on a 10 million node grid could require about 1 day CPU time on the Cray 2 machine.

Finally Section 5.4 was devoted to the development and analysis of a nonlinear system solver for the case of unsaturated flow. The nonlinear SIP solver was developed by adding an iterative linearization scheme to the previous iterative matrix solver (nested Picard iterations). A preliminary analysis of numerical requirements suggested that, in the case of transient infiltration on dry soils, there could be a severe limitation on the time step size in order to ensure the convergence of the nonlinear iteration loop (outer iterations). On the other hand, the SIP matrix solver is likely to converge much faster for any given time step of a transient problem (particularly for small time steps) than for the single step of a steady state problem. Our discussion of these issues also included a brief literature review. Our attempt at elucidating the space-time resolution requirements by way of numerical analysis was not entirely successful. The most remarkable finding in that study was perhaps the *grid Peclet number constraint* ($Pe = \alpha \Delta x_1 \leq 2$), which was obtained from a heuristic "nonlinear stability analysis" of the unsaturated finite difference system. This analysis also suggested that a very stringent requirement on the time step could result if the Peclet number constraint was not satisfied locally. The discussion included a physical interpretation of the Peclet number for unsaturated flow (gravity/diffusion), and focused on the possible divergence of the nonlinear solver in

severe cases such as infiltration in *dry heterogeneous* soils (sharp fronts).

It was felt that the complexity of the unsaturated flow problem, and the associated numerical issues, required a careful testing of the unsaturated flow simulator. The last part of Section 5.4 was devoted to numerical tests with increasingly complex problems of transient infiltration. These numerical experiments included specialized tests for checking certain features of the code such as the variable time step and the variable domain procedures, mass balance tests, comparisons with analytical solutions, and a number of infiltration experiments with homogeneous and uniformly layered soils, mostly in two dimensions. Our overall conclusion from these numerical experiments is that the unsaturated flow simulator appears as a reliable and flexible tool for the detailed simulation of fairly complex transient infiltration problems.

The encouraging results obtained here need however to be confirmed for the case of truly large realizations of random soil systems. This will be the subject of Chapter 7, where we will analyze in detail the solutions obtained from large single-realization simulations of flow in random soils, with three-dimensional grids on the order of ten thousand to several hundred thousand nodes. On the other hand, the forthcoming

Chapter 6 will be devoted to a systematic statistical analysis of the solutions obtained from large single-realization simulations of saturated flow. Note that both Chapter 6 and Chapter 7 will focus mostly on physical interpretation and mathematical analysis of numerical solutions, and only occasionally on some numerical issues. It is assumed at this stage that the flow simulator has been fully tested, both in the saturated and unsaturated flow regimes.