

A NUMERICAL EIGENVALUE STUDY OF PRECONDITIONED NON-EQUILIBRIUM TRANSPORT EQUATIONS

GIUSEPPE GAMBOLATI* AND GIORGIO PINI

Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, University of Padua, Padua, Italy

SUMMARY

The finite element integration of non-equilibrium contaminant transport in porous media yields sparse, unsymmetric, real or complex equations, which may be solved by iterative projection methods, such as Bi-CGSTAB and TFQMR, on condition that they are effectively preconditioned. To ensure a fast convergence, the eigenspectrum of the preconditioned equations has to be very compact around unity. Compactness is generally measured by the spectral condition number. In difficult advection-dominated problems, however, the condition number may be large and nevertheless, convergence may be good. A numerical study of the preconditioned eigenspectrum of a representative test case is performed using the incomplete triangular factorization. The results show that preconditioning eliminates most of the original complex eigenvalues, and that compactness is not necessarily jeopardized by a large condition number. Quite surprisingly, it is shown that the preconditioned complex problem may have a more compact real eigenspectrum than the equivalent real problem. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: non-equilibrium transport; finite elements; non-symmetric eigenproblem; projection methods

1. INTRODUCTION

Recently, new finite element models have been developed to solve the problem of non-equilibrium or non-ideal reactive transport in sorbing porous media [1–6]. These models may account for chemical diffusion, mechanical dispersion, advection, decay, instantaneous and rate-limited sorption onto the solid phase, and differ from each other for the numerical approach of the solution. The resulting algebraic equations involve large sparse matrices that are generally unsymmetric and may be either real or complex [7,8]. Solution is obtained with the aid of iterative projection (or conjugate gradient-like) methods, effectively preconditioned [9]. In particular, the present study focuses on the two solvers known as Bi-CGSTAB [10] and TFQMR [11].

The asymptotic convergence rate of these solvers is of the greatest importance for an efficient and robust use of the related models in both field and parametric analyses. Convergence depends on a large number of factors, including the physico-chemical properties of fluid, soil and contaminant, on the type of elements and structure of the mesh, and can hardly be described by a few theoretical numbers. It is generally recognized, however, that precondition-

* Correspondence to: Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, University of Padua, Padua, Italy.

ing is essential to ensure convergence, and that a good preconditioner should compact the eigenspectrum around unity, while requiring a limited computational effort so as to maintain the iterative solver competitive with other alternative solution techniques.

A rough relatively inexpensive measure of the eigenspectrum compactness is provided by the spectral condition number. In difficult convection-dominated problems, where the skew symmetric component is important, the condition number may not be indicative of the spectrum compactness, and hence may not give helpful information on the rate of convergence of the projection solver.

The present paper describes some interesting numerical findings concerned with the distribution of the real and complex eigenvalues before and after preconditioning in a medium-large finite element sample problem of non-equilibrium transport in a partially saturated porous system. Three solution approaches, referred to as coupled, decoupled and FELT (Finite Element Laplace Transform) [7], are considered and discussed. The influence of high Peclet and Courant numbers is also examined. The real, as well as the complex FELT equations are preconditioned by the incomplete triangular factors with no fill-in (ILU(0) [12,13]) and partial controlled fill-in (ILUT [14]), as there is much literature, e.g. [15], showing that the preconditioners belonging to the class of incomplete factorization are quite robust preconditioners for a large set of numerical problems. The full eigenspectrum of the native equations is computed and compared with that after preconditioning. The frequency of the complex eigenvalues in both the original and the preconditioned matrix is obtained, and is shown to be highly influenced by the strength of the advective component. Preconditioning may significantly reduce the number of complex eigenvalues but does not eliminate them all. Surprisingly enough, the fully complex FELT equations turned out to possess an almost entirely real eigenspectrum after preconditioning, thus providing indirect evidence of the excellent performance of the projection solvers in the complex field.

The paper is organized as follows. First a brief review of solution methods to the non-equilibrium transport model is provided. Then the numerical example is introduced and the eigenspectrums of the native and preconditioned matrices are computed, compared and discussed. A few representative convergence profiles are also shown. Finally, a set of concluding remarks are issued.

2. REVIEW OF NUMERICAL SOLUTION APPROACHES FOR THE NON-EQUILIBRIUM TRANSPORT MODEL

The conceptualization of van Genuchten and Wierenga [16] is used for the first-order kinetics representation of the dual-porosity model. This model describes non-equilibrium contaminant transport in a variably saturated, aggregated porous medium, where the saturated pore space is subdivided into a mobile water region and an immobile water region. Fluid flow and convective and dispersive solute transport occur in the mobile region only, and the exchange of solute between the mobile and immobile regions is controlled by a diffusive mechanism. The model is further enhanced by introducing linear equilibrium sorption and a biodegradation or radioactive decay term in both the mobile and immobile regions. Under these assumptions, the general equations describing the linear dual-porosity model in a two-dimensional system are [2]:

$$\frac{\partial}{\partial x_i} \left(D_{ij} \frac{\partial c_m}{\partial x_j} \right) - v_i \frac{\partial c_m}{\partial x_i} = T_m \frac{\partial c_m}{\partial t} + T_{im} \frac{\partial c_{im}}{\partial t} + \lambda (T_m c_m + T_{im} c_{im}) + q(c_m - c^*) - f; \quad (1a)$$

$$i, j = 1, 2,$$

$$T_{im} \frac{\partial c_{im}}{\partial t} = \alpha (c_m - c_{im}) - \lambda T_{im} c_{im}, \quad (1b)$$

where x_i is the i th co-ordinate direction; t is time; c_m and c_{im} are the concentrations of the dissolved constituent in the mobile and immobile water regions respectively; the dispersion coefficient is $D_{ij} = \alpha_T |v| \delta_{ij} + (\alpha_L - \alpha_T) v_i v_j / |v| + n_m S_w D_0 \tau \delta_{ij}$; α_L and α_T are the longitudinal and transverse dispersivities respectively; v_i is the Darcy velocity; $|v| = \sqrt{v_1^2 + v_2^2}$; δ_{ij} is the Kronecker delta; D_0 is the molecular diffusion coefficient; τ is the tortuosity; n_m and n_{im} are the porosities of the mobile and immobile regions; S_w is the water saturation in the mobile region; $T_m = n_m S_w + \rho_s F k_{d_m}$ and $T_{im} = n_{im} + \rho_s (1 - F) k_{d_{im}}$ are retardation factors for the mobile and immobile zones; $\rho_s = (1 - n_m - n_{im}) \gamma_s$ is the bulk soil density; γ_s is the density of the solid grains; F is the fraction of sorption sites in direct contact with the mobile water; k_{d_m} and $k_{d_{im}}$ are the distribution coefficients in the linear Freundlich isotherm describing instantaneous sorption in the mobile and immobile regions; λ is the linear decay constant; q represents distributed source/sink terms; f is the distributed flow rate of the solute per unit volume; c^* is the concentration of the injected/withdrawn fluid; and α is the mass transfer coefficient for the diffusion process between the mobile and immobile water regions. Parameter α provides an indication of how close to equilibrium the system is. As $\alpha \rightarrow \infty$, the mass exchange between the mobile and the immobile regions becomes instantaneous, and the transport model reduces to the classical convection–dispersion–reaction equation with the coefficient multiplying $\partial c_m / \partial t$ equal to $T_m + T_{im}$. Likewise, if $\alpha \rightarrow 0$, $\partial c_{im} / \partial t = -\lambda c_{im}$, implying that $c_{im} = 0$ everywhere (for initial condition $c_{im} = 0$), and hence, the convection–dispersion–reaction equation holds true again with $T_m + T_{im}$ replaced by T_m [1]. The velocity field represented by vector v and the saturation values represented by S_w can be obtained by solving the flow equation in variably saturated porous media for the mobile region [4].

Using (1b), (1a) can be written as

$$\frac{\partial}{\partial x_i} \left(D_{ij} \frac{\partial c_m}{\partial x_j} \right) - v_i \frac{\partial c_m}{\partial x_i} = T_m \frac{\partial c_m}{\partial t} + \alpha (c_m - c_{im}) + \lambda T_m c_m + q(c_m - c^*) - f, \quad i, j = 1, 2. \quad (2)$$

Equations (2) and (1b) are both explicitly used in the coupled approach to the solution of the dual-porosity model.

Integrating Equation (1b) analytically, assuming $c_{im}(x_i, t = 0) = 0$, and substituting the result into (1a) leads to an integro-differential equation for the mobile region concentration [2]:

$$\begin{aligned} \frac{\partial}{\partial x_i} \left(D_{ij} \frac{\partial c_m}{\partial x_j} \right) - v_i \frac{\partial c_m}{\partial x_i} \\ = T_m \frac{\partial c_m}{\partial t} + (\alpha + \lambda T_m + q) c_m - (q c^* + f) - \alpha \beta e^{-(\beta + \lambda)t} \int_0^t e^{(\beta + \lambda)\tau} c_m(\tau) d\tau, \quad i, j = 1, 2, \end{aligned} \quad (3)$$

where $\beta = \alpha / T_{im}$. Equation (3) forms the basis for the integro–differential approach to the solution of the dual-porosity model.

2.1. Coupled approach

Equations (2) and (1b) are integrated by linear triangular finite elements in space and finite differences in time using the Galerkin formulation and a weighted time stepping scheme [7]. Denoting \mathbf{c}_m and \mathbf{c}_{im} as the vectors containing the unknown mobile and immobile concentrations at each of the N nodes of the finite element mesh, the following algebraic system is obtained:

$$\begin{aligned} & \left[v_1(\mathbf{S} + \mathbf{B} + \tilde{\mathbf{E}} + \tilde{\mathbf{F}})^{k+v_1} + \frac{1}{\Delta t^k} \tilde{\mathbf{G}}^{k+v_1} \right] \mathbf{c}_m^{k+1} - v_1 \mathbf{R} \mathbf{c}_{im}^{k+1} \\ & = \left[\frac{1}{\Delta t^k} \tilde{\mathbf{G}}^{k+v_1} - v_{11}(\mathbf{S} + \mathbf{B} + \tilde{\mathbf{E}} + \tilde{\mathbf{F}})^{k+v_1} \right] \mathbf{c}_m^k + v_{11} \mathbf{R} \mathbf{c}_{im}^k - \mathbf{r}^{*,k+v_1}, \end{aligned} \quad (4a)$$

$$\left[\frac{1}{\Delta t^k} \mathbf{G}^* + v_2 \mathbf{R}^* \right] \mathbf{c}_{im}^{k+1} - v_2 \mathbf{R} \mathbf{c}_m^{k+1} = \left[\frac{1}{\Delta t^k} \mathbf{G}^* - v_{22} \mathbf{R}^* \right] \mathbf{c}_{im}^k + v_{22} \mathbf{R} \mathbf{c}_m^k, \quad (4b)$$

where v_1 and v_2 ($0 < v_1, v_2 \leq 1$) are weighting factors; $v_{11} = 1 - v_1$; $v_{22} = 1 - v_2$; k indicates the time level; \mathbf{S} , \mathbf{B} , and $\tilde{\mathbf{G}}$ are the stiffness, advection and capacitance matrices respectively; $\tilde{\mathbf{E}}$, $\tilde{\mathbf{F}}$, \mathbf{R} and \mathbf{R}^* are capacitance-type matrices arising from the \mathbf{c}_m and \mathbf{c}_{im} terms on the right-hand-side of Equations (2) and (1b), and from the convective component of the Cauchy boundary conditions; \mathbf{G}^* is the capacitance matrix for Equation (1b); and \mathbf{r}^* contains source/sink terms, Neumann boundary conditions, and the total solute flux across the Cauchy boundary. Equations (4a) and (4b) represent a non-symmetric system of $2N$ coupled equations for the nodal mobile and immobile region concentrations.

2.2. Decoupled approach

Following Leismann *et al.* [17], a weighted time difference approximation was firstly applied to the coupled Equations (2) and (1b). Obtained for (1b) was [1]:

$$\mathbf{c}_{im}^{k+1} = \frac{T_{im} \mathbf{c}_{im}^k + \alpha \Delta t^k [v_2 \mathbf{c}_m^{k+1} + v_{22} \mathbf{c}_m^k] - \Delta t^k (\alpha + \lambda T_{im}) v_{22} \mathbf{c}_{im}^k}{v_2 \Delta t^k (\alpha + \lambda T_{im}) + T_{im}}, \quad (5)$$

which is substituted for \mathbf{c}_{im}^{k+1} in the discretized form of (2). Integrating the equation in space by the Galerkin method gives a decoupled system in the N unknown mobile region concentrations \mathbf{c}_m^{k+1} :

$$\begin{aligned} & \left[v_1(\mathbf{S} + \mathbf{B} + \tilde{\mathbf{E}} + \tilde{\mathbf{F}})^{k+v_1} + \frac{1}{\Delta t^k} \tilde{\mathbf{G}}^{k+v_1} - v_2 \mathbf{E}^* \right] \mathbf{c}_m^{k+1} \\ & = \left[\frac{1}{\Delta t^k} \tilde{\mathbf{G}}^{k+v_1} - v_{11}(\mathbf{S} + \mathbf{B} + \tilde{\mathbf{E}} + \tilde{\mathbf{F}})^{k+v_1} - v_{22} \mathbf{E}^* \right] \mathbf{c}_m^k + \mathbf{E}^{**} \mathbf{c}_{im}^k - \mathbf{r}^{*,k+v_1}, \end{aligned} \quad (6)$$

where \mathbf{E}^* and \mathbf{E}^{**} are capacitance-type matrices involving the terms $v_1 \alpha^2 \Delta t^k / \delta$ and $\alpha [T_{im} + \Delta t^k (\alpha + \lambda T_{im}) (v_2 - v_1)] / \delta$ respectively, where $\delta = v_2 \Delta t^k (\alpha + \lambda T_{im}) + T_{im}$.

2.3. FELT approach

Provided that the non-equilibrium transport model is used in its simplified linear form, the FELT approach can be used for the discretization of Equation (3) [18].

Using this technique, Equation (3) is first transformed into the Laplace domain. To this aim, let \mathcal{L} be the Laplace transformation operator, defined as:

$$\mathcal{L}[c_m(t)] = \bar{c}_m(p) = \int_0^{+\infty} e^{-pt} c_m(t) dt,$$

where p is the Laplace transform parameter ($p \in \mathbb{C}$). Recall that the first fundamental property of the Laplace transformation is:

$$\mathcal{L}\left[\frac{\partial c_m}{\partial t}\right] = p\bar{c}_m - c_m(x_i, t=0)$$

and that a convolution integral is transformed into a product in the Laplace domain:

$$\mathcal{L}[c_m * g] = \mathcal{L}\left[\int_0^t c_m(\tau)g(t-\tau) d\tau\right] = \mathcal{L}[c_m]\mathcal{L}[g] = \bar{c}_m\bar{g}.$$

Noting that:

$$\mathcal{L}[e^{-(\beta+\lambda)t}] = \frac{1}{p + \beta + \lambda},$$

the Laplace transformed transport equation (on condition that the parameters are time-independent) can be written as:

$$\frac{\partial}{\partial x_i} \left(D_{ij} \frac{\partial \bar{c}_m}{\partial x_j} \right) - v_i \frac{\partial \bar{c}_m}{\partial x_i} = T_m(p\bar{c}_m - c_m(x_i, t=0)) + (\alpha + \lambda T_m + q)\bar{c}_m - (q\bar{c}^* + \bar{f}) - \frac{\alpha\beta\bar{c}_m}{p + \beta + \lambda}$$

or, after rearrangement of the terms:

$$\frac{\partial}{\partial x_i} \left(D_{ij} \frac{\partial \bar{c}_m}{\partial x_j} \right) - v_i \frac{\partial \bar{c}_m}{\partial x_i} = \left[T_m(p + \lambda) + q + \frac{\alpha T_m(p + \lambda)}{T_{im}(p + \lambda) + \alpha} \right] \bar{c}_m - q\bar{c}^* - \bar{f} - T_m c_m(x_i, t=0). \quad (7)$$

Boundary conditions to (7) are also properly transformed into the Laplace domain:

$$\bar{c}_m(x_i, p) = \bar{c}_1(x_i, p) \quad \text{Dirichlet b.c.}$$

$$D_{ij} \frac{\partial \bar{c}_m}{\partial x_j} n_i = \bar{q}_c^D(x_i, p) \quad \text{Neumann b.c.}$$

$$\left(D_{ij} \frac{\partial \bar{c}_m}{\partial x_j} - v_i \bar{c}_m \right) n_i = \bar{q}_c^T(x_i, p) \quad \text{Cauchy b.c.}$$

Equation (7) together with the transformed boundary conditions represent a PDE in the Laplace domain, where time has been removed. The solution of this equation gives the concentration values in the p space, or Laplace domain.

The FELT technique proceeds by solving, via the finite element scheme, Equation (7) for an appropriate number of p values. A set of solution vectors in the so-called p space is thus obtained. These vectors have to be transformed back into the time domain to provide the final mobile region concentration as a function of time.

In this paper, the numerical discretization of (7) for each value of p is obtained using a Galerkin approach with triangular elements and linear basis functions. The transformed concentration \bar{c}_m is approximated by:

$$\bar{c}_m \approx \hat{c} = \sum_{j=1}^N \bar{c}_j W_j(x_1, x_2), \quad (8)$$

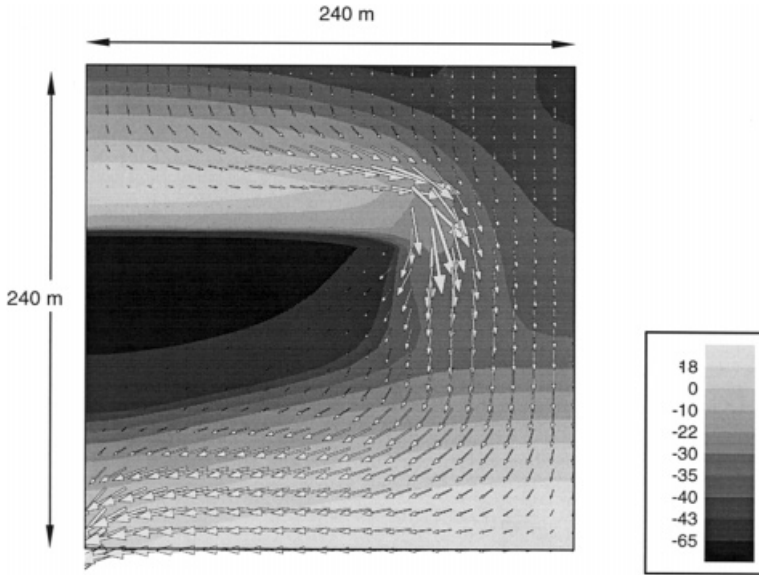


Figure 1. Pressure head (m) and velocity field (m day^{-1}) [9].

where $\bar{c} = (\bar{c}_1, \dots, \bar{c}_N)^T$ is the complex-valued vector of nodal concentrations in the transformed p space. Substituting (8) into (7), imposing the orthogonality condition between the residual and the test functions W_j , and applying Green's lemma to the dispersive terms only [19], leads to the final set of $N \times N$ algebraic equations [1]:

$$[\mathbf{S} + \mathbf{B} + \mathbf{E} + \mathbf{F}]\bar{c} + \mathbf{d} = 0. \quad (9)$$

The real matrices \mathbf{S} , \mathbf{B} and \mathbf{E} are the standard finite element matrices discretizing the dispersive, advective and Cauchy boundary condition components of Equation (7). Matrix \mathbf{F} is a complex-valued capacitance-type matrix arising from the discretizations of the terms involving the Laplace parameter p . Finally, the complex-valued vector \mathbf{d} implements the transformed boundary conditions together with the source and sink terms. The system (Equation (9)) is a set of linear algebraic equations in the complex space, and is a function of the Laplace parameter p . The solution to (9) provides the concentration in the Laplace domain. Anti-transformation of \bar{c} from the Laplace domain to the time domain gives the solution of Equation (3) as a function of time.

The numerical anti-transformation of \bar{c} is obtained by means of the *epsilon* algorithm [20], with the help of the refinement technique proposed in [21]. By the *epsilon* algorithm, a discrete set of values of the Laplace parameter p is chosen:

Table I. Description of parameters of test case considered

Case	Approach	Pe	Cr	α (day^{-1})	Δt (day)	α_L (m)
a1	Decoupled	1	1	0.03	5	2.5
a2	Decoupled	10	10	0.75	50	0.25
b1	Coupled	1	1	0.03	5	2.5
b2	Coupled	10	10	0.75	50	0.25
c1	FELT	1	1	0.03	5	2.5
c2	FELT	10	10	0.75	50	0.25

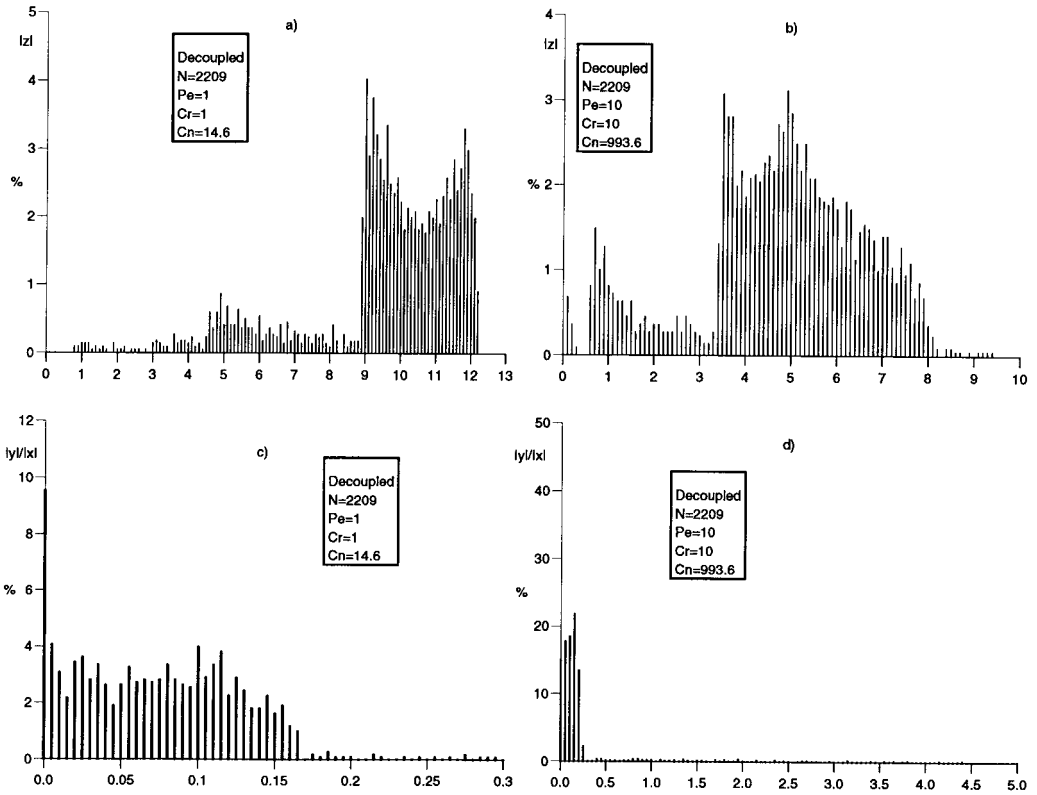


Figure 2. (a) Histogram of the absolute eigenvalue distribution of the coefficient matrix A for problem a1, the size of the classes is 0.1; (b) the same as (a) for problem a2; (c) histogram of the ratio between the imaginary and the real part of the eigenvalues of the coefficient matrix A for problem a1, the size of the classes is 0.005; (d) the same as (c) for problem a2 with the size of the classes equal to 0.05.

$$p = p_k = p_0 + ik\pi/T, \quad k = 0, 1, 2, \dots, 2M + 1, \tag{10}$$

where $p_0 = -\ln(\epsilon)/(1.6t_{\max})$, ϵ being the absolute error term and t_{\max} the maximum simulation time, and $T = 0.8t_{\max}$. With the choice of p as in (10), the system (Equation (9)) has to be solved for $r = 2(M + 1)$ values of the discrete Laplace parameter p_k .

The anti-transformed concentration $c_{j,k}$ at node j for $p = p_k$ is approximated by:

$$c_j(t) \approx \frac{e^{p_0 t}}{T} \left\{ \frac{1}{2} \bar{c}_{j,0} + \sum_{k=1}^{2M+1} \left[\text{Re}(\bar{c}_{j,k}) \cos\left(\frac{k\pi t}{T}\right) - \text{Im}(\bar{c}_{j,k}) \sin\left(\frac{k\pi t}{T}\right) \right] \right\}. \tag{11}$$

Denoting the inverse Laplace transformation by \mathcal{L}^{-1} , the approximate concentration in the time domain can then be written as:

$$\hat{c}(x_1, x_2, t) = \sum_{j=1}^N \mathcal{L}^{-1}[\bar{c}_j(p_k)] W_j(x_1, x_2) = \sum_{j=1}^N c_j(t) W_j(x_1, x_2), \tag{12}$$

where the real-valued vector $c(t) = (c_1(t), \dots, c_N(t))^T$ is the vector of nodal mobile concentrations at time t .

The value of M has to be set large enough to guarantee convergence of the series on the right-hand-side of Equation (11) to the correct concentration $c_j(t)$. In practice, M is chosen in

the range 5–40, since round-off errors may become dominant for larger M [21]. The value of M is also influenced by the presence of steep gradients in the solution (e.g. advective transport). The parameter t_{\max} has to be chosen based on the simulation time. The computation of the solution at small times relative to t_{\max} ($t \ll t_{\max}$), requires, in principle, a large value of M , but this may trigger round-off errors in the inversion algorithm. In these cases it may be necessary to solve the transport problem for a few t_{\max} values, which together, cover the prescribed maximum simulation time.

3. PROJECTION METHODS FOR THE SOLUTION TO FINITE ELEMENT EQUATIONS

Now write the final set of algebraic finite element equations in the general form

$$\mathbf{A}c = \mathbf{b}, \quad (13)$$

where \mathbf{A} , the coefficient matrix of Equations (4), (6) and (9), is sparse, non-symmetric, real or complex. Let N be the size of \mathbf{A} , keeping in mind that in the coupled approach, the system is

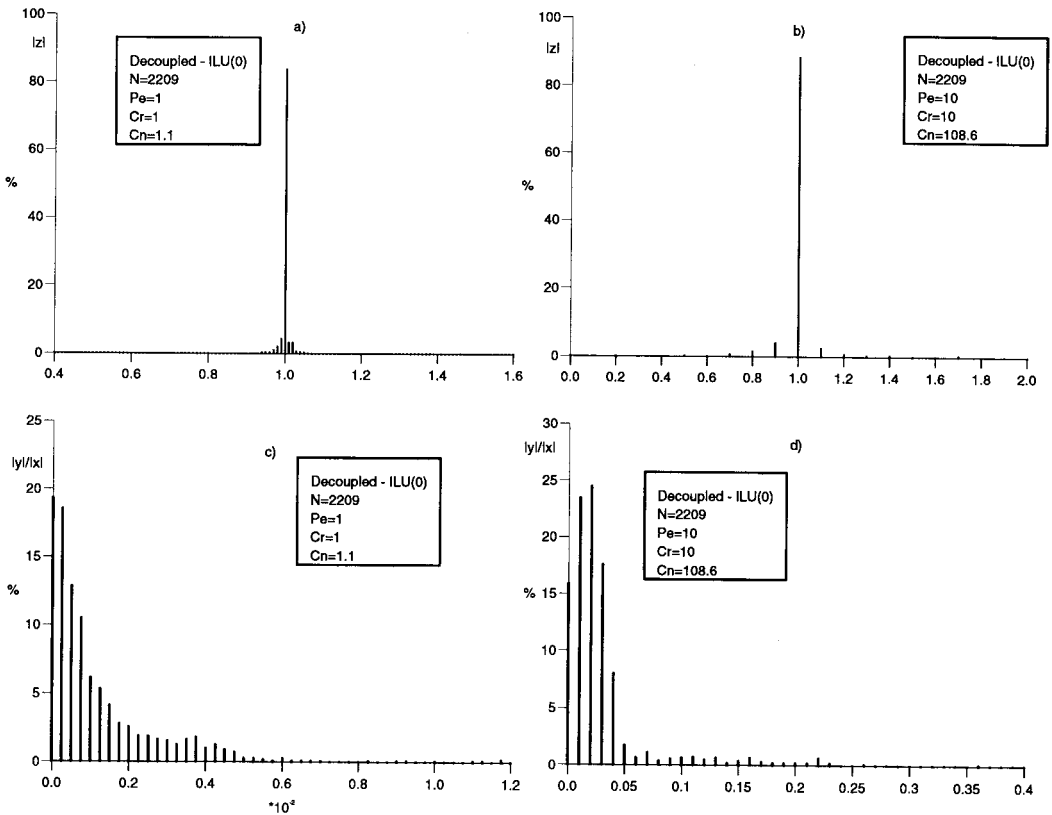


Figure 3. (a) Histogram of the absolute eigenvalue distribution of the preconditioned matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ (ILU(0)) for problem a1, the size of the classes is 0.01; (b) the same as (a) for problem a2 with the size of the classes equal to 0.1; (c) histogram of the ratio between the imaginary and the real part of the eigenvalues of the preconditioned matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ (ILU(0)) for problem a1, the size of the classes is $0.125 \cdot 10^{-3}$; (d) the same as (c) for problem a2 with the size of the classes equal to 0.01.

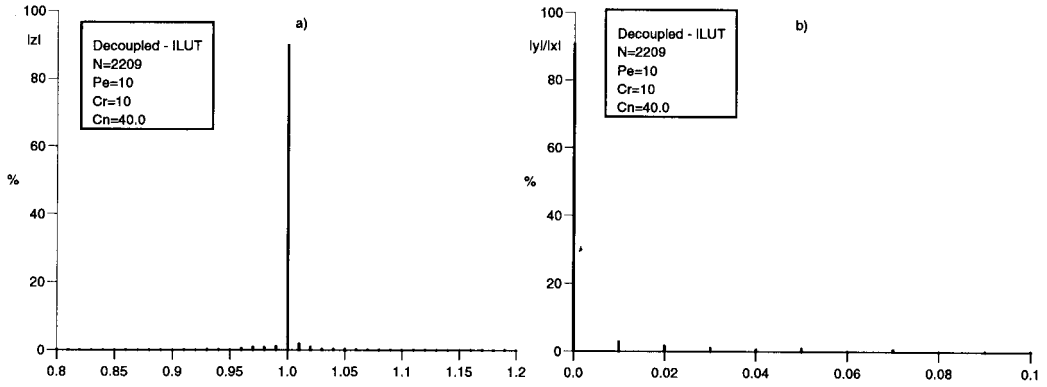


Figure 4. (a) Histogram of the absolute eigenvalue distribution of the preconditioned matrix $L^{-1}AU^{-1}$ (ILUT) for problem a2, the size of the classes is 0.01; (b) histogram of the ratio between the imaginary and the real part of the eigenvalues of the preconditioned matrix $L^{-1}AU^{-1}$ (ILUT) for problem a2, the size of the classes is 0.01.

twice as large as in the decoupled and FELT approaches. In recent years, projection (or conjugate gradient-like) solution methods have been widely and successfully used. These methods project $A\mathbf{c} = \mathbf{b}$ onto subspaces (called Krylov subspaces) of increasing size ℓ , and solve the projected system. This procedure has the remarkable property of terminating at the desired solution after a finite number of iterations in exact arithmetic, and in this respect, a projection method can be regarded as a direct approach. In practice, convergence long before the dimension of the subspace reaches its maximum possible value (N) is sought. On the other hand, round-off errors can prevent convergence to the desired solution within a finite number of iterations, so that in effect, projection methods are regarded as iterative methods. For a thorough review of projection methods see [13]. Bi-CGSTAB [10] and TFQMR [11] prove to be quite robust and efficient conjugate gradient-like methods for the transport problems addressed by the present study [9]. They are both variants of the BCG (biconjugate gradient) method [22,23], with the aim to avoid the explicit multiplication between A^T and a vector (while preserving the theoretical properties of BCG), and to smooth the erratic residual behavior of the CGS (conjugate gradient squared) method, the earliest BCG variant that did not require the direct use of the transpose of A [24].

The projection methods used in this paper do not require the use of optimal acceleration parameters. However, in practical problems of realistically large size, it is crucial to combine Krylov subspace algorithms with an effective preconditioning technique, which can prove to be the key factor for the success of any CG-like scheme.

The basic idea of preconditioning is as follows. Let A_1A_2 be a given non-singular $N \times N$ matrix, which in some measure approximates the coefficient matrix A of the discrete finite element system. This system is replaced by the equivalent system

$$A_1^{-1}AA_2^{-1}A_2\mathbf{c} = A_1^{-1}\mathbf{b}, \quad (14)$$

which can be written as

$$A'\mathbf{c}' = \mathbf{b}', \quad (15)$$

where $A' = A_1^{-1}AA_2^{-1}$, $\mathbf{c}' = A_2\mathbf{c}$ and $\mathbf{b}' = A_1^{-1}\mathbf{b}$. System (15) is now solved by Bi-CGSTAB or TFQMR in place of Equation (13), with \mathbf{c} recovered as $\mathbf{c} = A_2^{-1}\mathbf{c}'$.

One of the most popular and cost-effective preconditioners is derived from the class of the incomplete triangular factorizations of \mathbf{A} . Here, set $\mathbf{A}_1 = \mathbf{L}$ and $\mathbf{A}_2 = \mathbf{U}$, where \mathbf{L} and \mathbf{U} are lower and upper triangular matrices, and a form is obtained which resembles that of the true triangular factors of \mathbf{A} .

In the so-called incomplete Crout factorization, referred to herein as ILU, \mathbf{L} and \mathbf{U} are obtained by performing the standard LU decomposition of \mathbf{A} and dropping all fill-in elements, which are generated during the process [12,13]. This is the most inexpensive incomplete factorization possible. Depending on the actual structure of \mathbf{A} , however, the matrix $(\mathbf{LU})^{-1}$ can be a poor approximation of \mathbf{A}^{-1} , resulting in poor acceleration of the native projection method. Hence, ILUT [14] can be used instead. The ILUT fill-in process is properly controlled by two parameters, τ' and ϵ' , related to the retention number of newly generated elements and their relative magnitude.

Note that in the FELT approach, matrix \mathbf{A} is a function of p_k , which implies the construction of a different preconditioner for each p_k value.

4. COMPUTATION OF REAL AND COMPLEX EIGENVALUES

For the eigenvalues of the equations derived in Section 3, the QR method [25] is used. It allows for the computation of all the real and complex eigenvalues of arbitrary unsymmetric matrices. More specifically, use has been made of routine DGEEV of LAPACK [26] for real matrices and CGEEV for complex-valued matrices. For a more detailed analysis of the methods implemented in the LAPACK package, see [27,28]. It has been possible to calculate all the eigenvalues in the largest ($N = 4610$) problem as well for both the native matrix \mathbf{A} and the preconditioned matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ on the Risc 600/390 with 288 RAM Mbyte. For larger

Table II. Effective eigenspectrum, outliers, effective spectral condition number C_n (excluding the outliers), number N_c of significant complex eigenvalues with $|y|/|x| \geq \delta$

Case	Preconditioner	Effective eigenspectrum	Outliers	C_n	N_c ($\delta = 0.01$)	N_c ($\delta = 0.05$)
a1	None	[0.83608E+0; 0.12198E+2]		14.59	1874	1366
a1	ILU(0)	[0.93839E+0; 0.10560E+1]		1.13	8	0
a1	ILUT	[0.99970E+0; 0.10004E+1]		1.00	0	0
a2	None	[0.94429E-2; 0.93827E+1]		993.62	2016	1677
a2	ILU(0)	[0.45309E-1; 0.49204E+1]	7432.2	108.60	1620	206
a2	ILUT	[0.51935E-1; 0.20783E+1]	56.419	40.02	166	60
b1	None	[0.43775E+0; 0.20154E+2]		46.04	3538	1934
b1	ILU(0)	[0.82917E+0; 0.11490E+1]		1.39	1814	40
b1	ILUT	[0.97975E+0; 0.10239E+1]		1.05	1152	0
b2	None	[0.28156E-2; 0.18787E+2]		6672.47	2136	2054
b2	ILU(0)	[0.16885E-1; 0.24962E+1]	23.618, 55.498, 4344.6	147.83	3660	1724
b2	ILUT	[0.28323E+0; 0.24204E+1]	5.1128	8.54	3000	1008
c1	None	[0.96151E+2; 0.36214E+3]		3.76	2209	2209
c1	ILU(0)	[0.92843E+0; 0.10632E+1]		1.14	673	0
c2	None	[0.89129E+1; 0.38500E+2]		4.32	2209	2209
c2	ILU(0)	[0.91311E+0; 0.10792E+1]		1.18	894	0

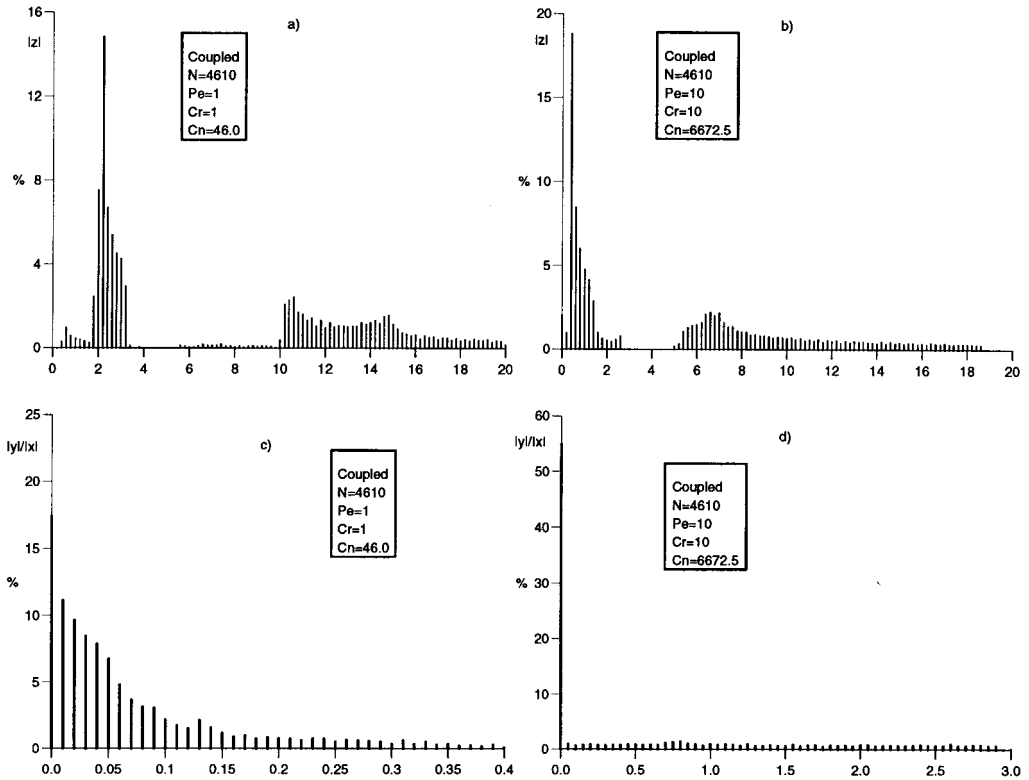


Figure 5. (a) Histogram of the absolute eigenvalue distribution of the coefficient matrix A for problem b1, the size of the classes is 0.2; (b) the same as (a) for problem b2; (c) histogram of the ratio between the imaginary and the real part of the eigenvalues of the coefficient matrix A for problem b1, the size of the classes is 0.01; (d) the same as (c) for problem b2 with the size of the classes equal to 0.05.

dimensions one could resort to the method known as ‘subspace iteration’ [29,30], which is suited for the assessment of the m ($m \ll N$) eigenvalues with the largest absolute module. The subspace iteration is a block generalization of the power method and is used in structural engineering [31]. Both techniques have been used for a cross-verification of the computed eigenspectrums.

5. DESCRIPTION OF TEST CASE

The test problem is a 240×240 m domain, uniformly discretized into 4608 triangles and 2401 nodes ($\Delta x = \Delta z = 5$ m). The steady state velocity and saturation values are obtained by solving the flow problem as mentioned earlier. For the flow problem, the domain is heterogeneous, with a horizontal slab of length 150 m and thickness 5 m starting at $x = 0$, $z = 165$ m. The slab has a saturated hydraulic conductivity 1000 times smaller than the rest of the domain. (The resulting steady state pressure head contours and velocity field are shown in Figure 1). The maximum velocity resulting from the flow problem is 1.09 m day $^{-1}$.

The transport problem is homogeneous, with $D_0 = c^* = f = k_{d_m} = \lambda = 0$, $\tau = 1.0$, $n_m = 0.30$, $n_{im} = 0.14$, $\gamma_s = 2700$ kg m $^{-3}$, $F = 0.4$, and $k_{d_{im}} = 0.5$ m 3 kg $^{-1}$. Dirichlet conditions of

$c_m = 1 \text{ kg m}^{-3}$ are imposed along the $x = 0$ and $z = 0$ boundaries, while $c_m = 0$ is imposed along $x = 240$ and $z = 240$. The initial condition is $c_m = c_{\text{im}} = 0$. The interest lies in seeing the effects of the mass transfer rate and of the Peclet (Pe) and Courant (Cr) numbers on the projection methods, where for this test problem $Pe = \Delta x / (2\alpha_L)$ and $Cr = v_{\text{max}} \Delta t / \Delta x \approx \Delta t / \Delta x$, so the model parameters that are varied are α , $\alpha_L (= \alpha_T)$ and Δt . The values of Pe , Cr and α used in the present numerical study are provided in Table I. The final problem size is $N = 2209$ ($N = 4610$ for the coupled example) to account for the equations that have been dropped on the Dirichlet boundary nodes. The non-zero coefficients of the corresponding finite element matrices are 15089 and 62432 respectively.

6. NUMERICAL RESULTS

In this section, the distribution of the eigenvalues $z = x + iy$ of the original matrix \mathbf{A} and the preconditioned one $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ are shown and commented on. Some representative convergence profiles are also given for the three solution approaches. Two types of representations are used, i.e. the frequency distribution of the absolute values $|z| = \sqrt{x^2 + y^2}$ and the ratios $|y|/|x|$ of the imaginary and real part of z . For the case of incomplete factorization with a

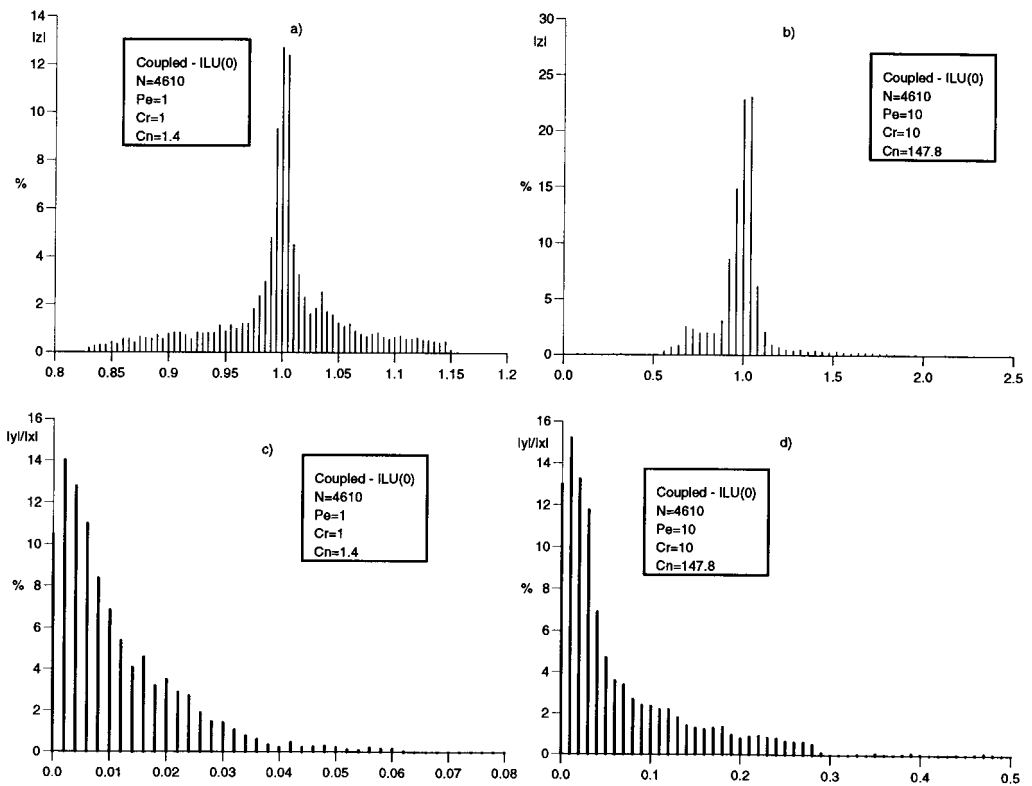


Figure 6. (a) Histogram of the absolute eigenvalue distribution of the preconditioned matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ (ILU(0)) for problem b1, the size of the classes is 0.005; (b) the same as (a) for problem b2 with the size of the classes equal to 0.04; (c) histogram of the ratio between the imaginary and the real part of the eigenvalues of the preconditioned matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ (ILU(0)) for problem b1, the size of the classes is 0.002; (d) the same as (c) for problem b2 with the size of the classes equal to 0.01.

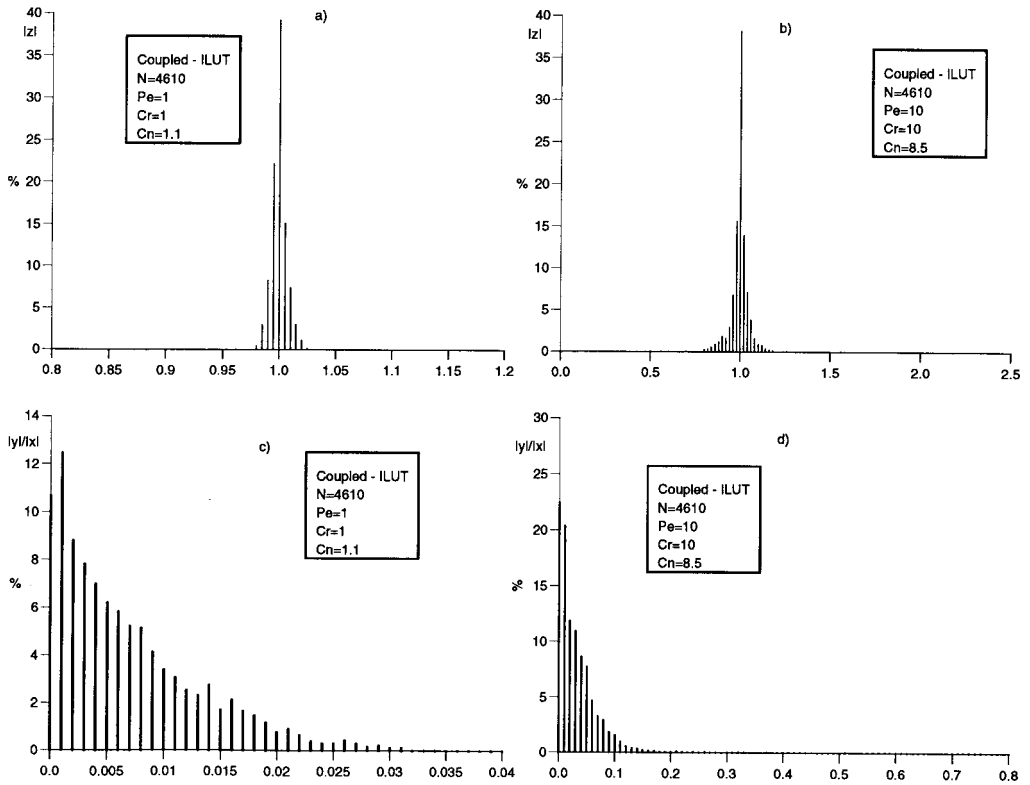


Figure 7. (a) Histogram of the absolute eigenvalue distribution of the preconditioned matrix $L^{-1}AU^{-1}$ (ILUT) for problem b1, the size of the classes is 0.005; (b) the same as (a) for problem b2 with the size of the classes equal to 0.02; (c) histogram of the ratio between the imaginary and the real part of the eigenvalues of the preconditioned matrix $L^{-1}AU^{-1}$ (ILUT) for problem b1, the size of the classes is 0.001; (d) the same as (c) for problem b2 with the size of the classes equal to 0.01.

controlled fill-in obtained with ILUT preconditioning, the parameters $\tau' = 5$, $\epsilon' = 10^{-9}$ have been assumed; for an explanation of this selection see [9]. For a correct interpretation of the following histograms, it must be kept in mind that the size of the histogram interval containing zero on the horizontal axis is half that of the subsequent classes.

With reference to Table I, Figure 2 shows the A eigenvalue distribution for problems a1 and a2, while the same results for the equations preconditioned with ILU(0) and ILUT are given in Figures 3 and 4 respectively. Note that in Figure 4, the results for the problem a1 are missing as ILUT compacts all the eigenvalues into only one class around unity. Inspection of the previous figures (see also Table II) shows that the interval that contains the eigenvalues increases for both A and $L^{-1}AU^{-1}$ when Pe and Cr increase, irrespective of preconditioning (ILU(0) or ILUT), and that the frequency of complex eigenvalues follows the same pattern. As is expected, ILUT provides a more compact eigenspectrum than ILU(0). Table II supplies the upper and lower bounds of the effective eigenspectrum with the outliers given separately, a measure $Cn = \max|z|/\min|z|$ of the effective condition number excluding the outliers, and the number N_c of representative complex characteristic values. N_c is formed by counting the eigenvalues for which $|y|/|x| > \delta$, with $\delta = 10^{-2}$ and $5 \cdot 10^{-2}$. Note that for problem a2 with high Pe and Cr , preconditioning produces an outlier, namely an eigenvalue falling outside the

effective eigenspectrum, which actually controls the solver asymptotic rate of convergence. Hence, for problem a2, which is convection-dominated, the estimate of the condition number formed with the true minimal and maximal eigenvalues of the preconditioned matrix is not indicative of the performance of Bi-CGSTAB and TFQMR. Actually, it may be observed that the estimated true condition number of $L^{-1}AU^{-1}$ is even higher than that of the unpreconditioned matrix A , for which both Bi-CGSTAB and TFQMR would converge much slower or would even fail to converge. Also, note in Table II the large number of complex eigenvalues in A , which is drastically reduced by preconditioning in problem a1, but much less in problem a2. The effective condition number Cn is still significantly larger than 1 in the more difficult problem a2, while it is very close to 1 for problem a1. The generation of outliers may be related to instability effects introduced by the incomplete triangular decomposition [32].

Figures 5–7 show results for problems b1 and b2 that are similar to those given in Figures 2–4; see also Table II. While a comparison of problems b1 and b2 leads to comments similar to the ones made before, a comparison of problems a and b indicates that the decoupled approach of solution (a) is better conditioned than the coupled one (b) with a lower frequency of complex eigenvalues. This also holds after preconditioning and suggests that coupling gives rise to a numerical problem that is more difficult to solve by Bi-CGSTAB and TFQMR, not to mention the number of equations, which is twice as big.

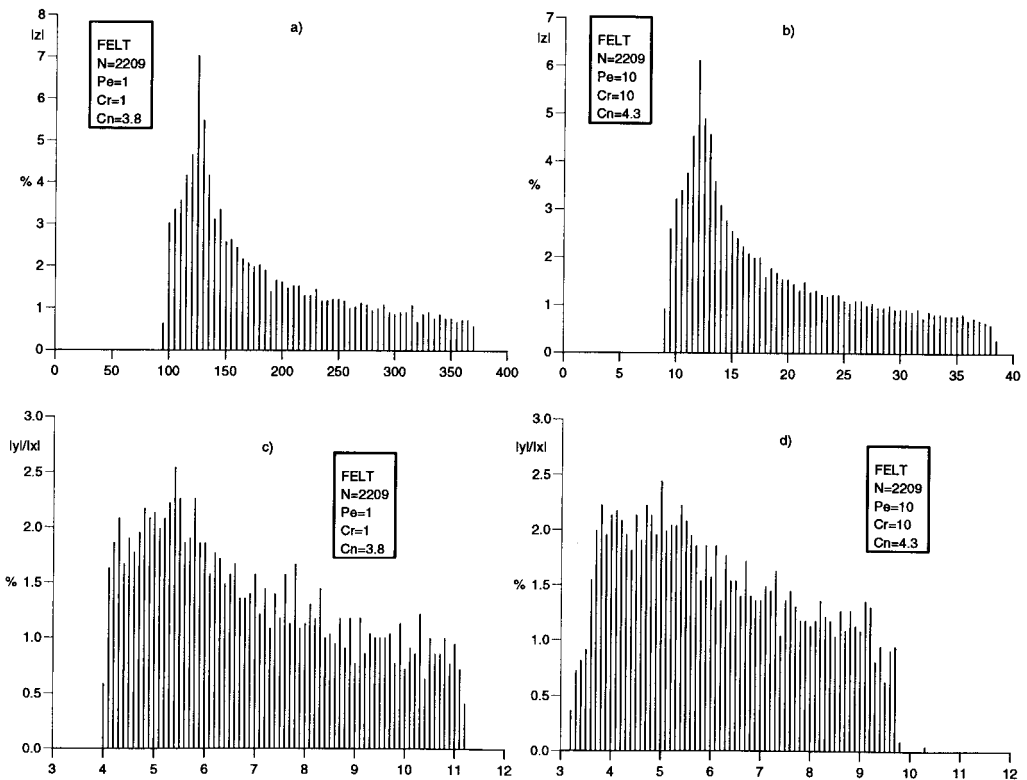


Figure 8. (a) Histogram of the absolute eigenvalue distribution of the coefficient matrix A for problem c1, the size of the classes is 5; (b) the same as (a) for problem c2 with the size of the classes equal to 0.5; (c) histogram of the ratio between the imaginary and the real part of the eigenvalues of the coefficient matrix A for problem c1, the size of the classes is 0.1; (d) the same as (c) for problem c2.

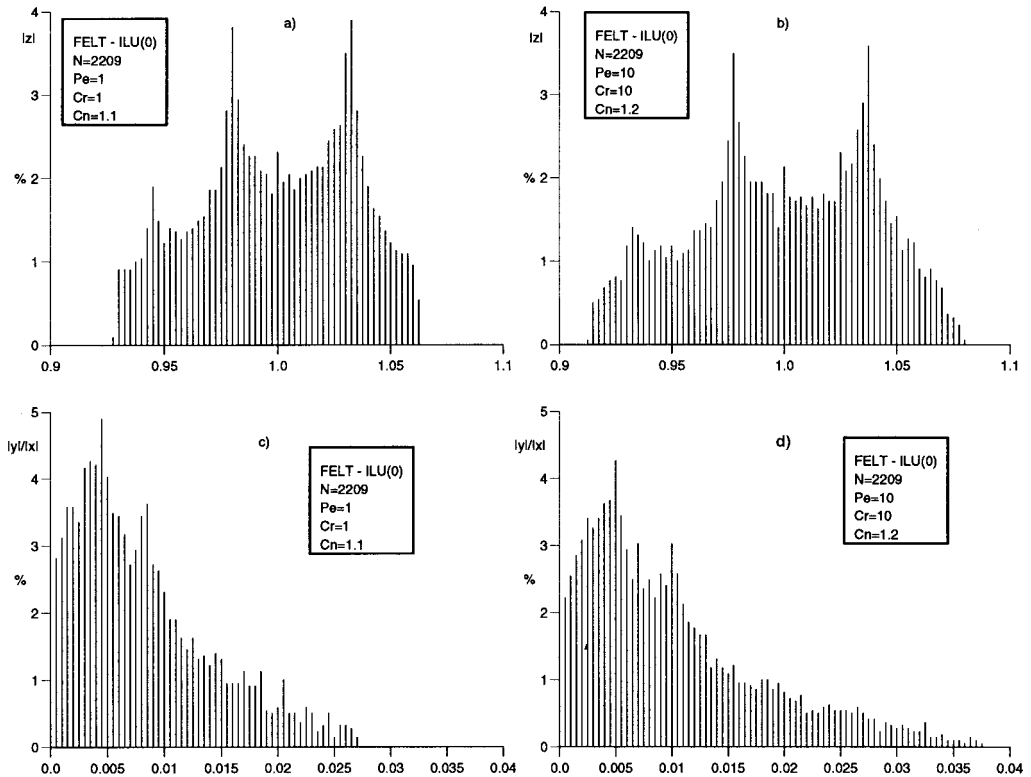


Figure 9. (a) Histogram of the absolute eigenvalue distribution of the preconditioned matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ (ILU(0)) for problem c1, the size of the classes is 0.0025; (b) the same as (a) for problem c2; (c) histogram of the ratio between the imaginary and the real part of the eigenvalues of the preconditioned matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ (ILU(0)) for problem c1, the size of the classes is $0.5 \cdot 10^{-3}$; (d) the same as (c) for problem c2.

Finally, the results from the FELT approach are provided in Figures 8 and 9 and in Table II. They have been obtained using a Laplace parameter p_k with $k = 63$ ($M = 31$), i.e. with the largest imaginary part, $\epsilon = 10^{-15}$ and $t_{\max} = 50$ day. Note that the complex \mathbf{A} condition number is much smaller than is for real \mathbf{A} , and preconditioning with ILUT is not reported since the ILUT preconditioned eigenspectrum is very compact around unity. Also, note the absence of outliers in problem c2 and the complete elimination from $\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1}$ of eigenvalues with an imaginary part 5% larger than the real part. The histograms of Figures 8 and 9 are more uniform than the equivalent histograms of Figures 2–7. The eigenanalysis show therefore, that Bi-CGSTAB and TFQMR are very well-suited to solve the FELT equations, consistent with the results obtained by Gambolati and Pini [33] in their convergence study of complex projection solvers.

As far as the convergence of Bi-CGSTAB and TFQMR is concerned, some representative profiles in terms of relative residual $\|r_r\|$ are given in Figures 10–12 for test case a2, b2 and c2 respectively, i.e. the most difficult problems from a numerical viewpoint. Note in Figures 10 and 11, the erratic behavior of Bi-CGSTAB and the slow convergence of TFQMR with the ILU(0) preconditioner. Using ILUT greatly improves the solver performance. Also note the FELT superior performance (Figure 12) in terms of stability and rate of convergence for both Bi-CGSTAB and TFQMR. This outcome is consistent with the eigenvalue distribution of the preconditioned complex approach, as discussed above.

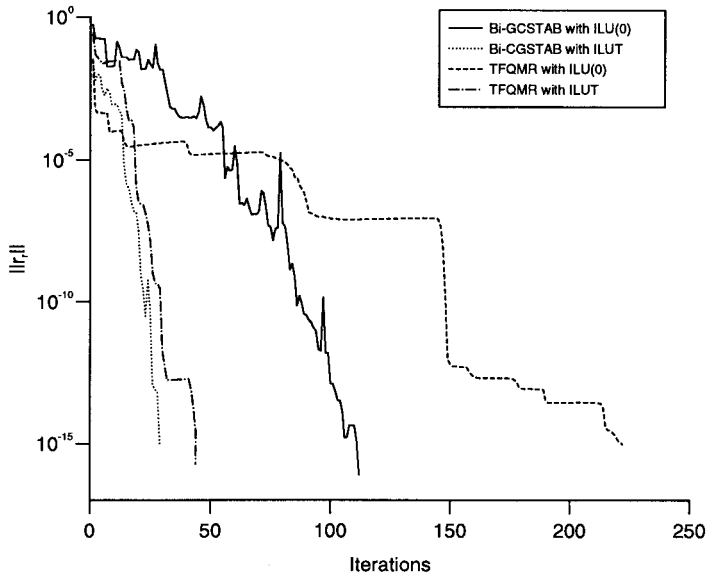


Figure 10. Convergence profiles of Bi-CGSTAB and TFQMR preconditioned with ILU(0) or ILUT for problem a2 (decoupled approach).

7. CONCLUSIONS

The following results are worth summarizing:

(1) The equations that arise from the finite element solution to the non-equilibrium transport problem in porous media may present a partially complex eigenspectrum with the

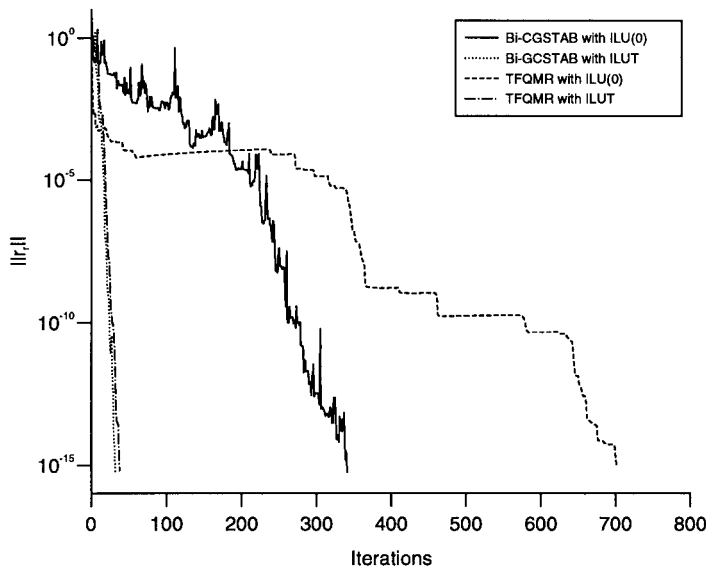


Figure 11. Convergence profiles of Bi-CGSTAB and TFQMR preconditioned with ILU(0) or ILUT for problem b2 (coupled approach).

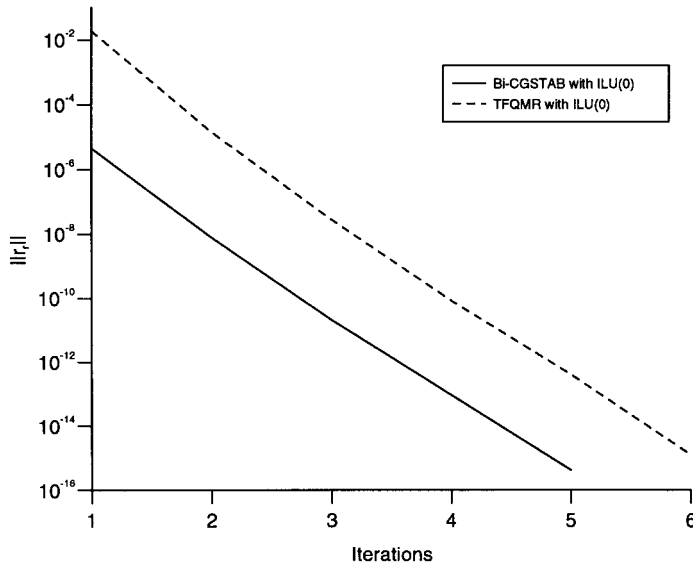


Figure 12. Convergence profiles of Bi-CGSTAB and TFQMR preconditioned with ILU(0) for problem c2 (FELT with $p = p_{63}$).

frequency of the complex eigenvalues increasing with increasing the Peclet and the Courant numbers.

(2) Solution by projection methods Bi-CGSTAB and TFQMR, which have proven quite robust in the problems addressed by the present study, are investigated by using a preconditioner based on the incomplete triangular factorization (either ILU(0) or ILUT) of the coefficient matrices.

(3) The eigenspectrum can still possess complex eigenvalues after preconditioning, especially in advection-dominated problems. However, a good preconditioner should lead to a drastic reduction of their frequency.

(4) The FELT preconditioned equations have an almost real eigenspectrum. This means that the incomplete triangular factorization is an excellent preconditioning technique for the complex solution approach as well.

(5) Bi-CGSTAB and TFQMR are more effectively preconditioned in the FELT formulation and, subordinately, in the decoupled approach of solution.

(6) The estimate of the spectral condition number formed with the smallest and largest eigenvalues may not be a good indicator of the eigenspectrum compactness, particularly when advection dominates and outliers may be easily generated by a numerically unstable incomplete factorization. Hence, it should not be used to provide an asymptotic evaluation of the solver convergence rate.

ACKNOWLEDGMENTS

This work has been supported in part by Fondi Ministeriali MURST 40% (cofinanziamento) and by INCO-COPERNICUS program of the European Community (Contract No. ER-BIC15CT960211). The authors would also like to thank two anonymous reviewers for their helpful and constructive comments.

REFERENCES

1. G. Gambolati, C. Gallo and C. Paniconi, 'Numerical integration methods for the dual-porosity model in sorbing porous media', in A. Peters, G. Wittum, B. Herrling, U. Meissner, C.A. Brebbia, W.G. Gray and G.F. Pinder (eds.), *Computational Methods in Water Resources X*, vol. 1, Kluwer, Dordrecht, 1994, pp. 621–628.
2. G. Gambolati, C. Paniconi and M. Putti, 'Mass transfer analysis in sorbing porous media by an integro-differential approach', in S.S.Y. Wang (ed.), *Advances in Hydro-Science and Engineering*, vol. I, part B, The University of Mississippi, University, MS, 1993, pp. 1819–1828.
3. G. Gambolati, G. Pini and M. Putti, 'Conjugate gradient-like methods for the numerical solution of the two site model in sorbing porous media', in S. Atluri, G. Yagawa, and T.A. Cruse (eds.), *Computational Mechanics 1995. Theory and Applications*, Springer, New York, 1995, pp. 748–753.
4. G. Gambolati, G. Pini, M. Putti and C. Paniconi, 'Finite element modeling of the transport of reactive contaminants in variably saturated soils with LEA and non-LEA sorption', in P. Zannetti (ed.), *Environmental Modeling. Vol. II: Computer Methods and Software for Simulating Environmental Pollution and its Adverse Effects*, Computational Mechanics Publications, Southampton, UK, 1994, pp. 173–212.
5. C. Paniconi, S. Ferraris, M. Putti, G. Pini and G. Gambolati, 'Three-dimensional numerical codes for simulating groundwater contamination: FLOW3D, flow in saturated and unsaturated porous media', in P. Zannetti (ed.), *Proc. Envirosoft 1994*, Computational Mechanics Publications, Southampton, UK, 1994, pp. 149–156.
6. M. Putti, S. Ferraris, C. Paniconi, G. Pini and G. Gambolati, 'Three-dimensional numerical codes for simulating groundwater contamination: TRAN3D, transport with equilibrium and non-equilibrium adsorption', in P. Zannetti (ed.), *Proc. Envirosoft 1994*, Computational Mechanics Publications, Southampton, UK, 1994, pp. 141–148.
7. C. Gallo, C. Paniconi and G. Gambolati, 'Comparison of solution approaches for the two-domain model of non-equilibrium transport in porous media', *Adv. Water Resour.*, **19**, 241–253 (1996).
8. G. Gambolati, C. Gallo, C. Paniconi and M. Putti, 'Numerical solutions for non-equilibrium solute transport in porous media', in *Advances in Hydro-Science and Engineering, Volume II, Part B*, Tsinghua University Press, Beijing, China, 1995, pp. 1733–1742.
9. G. Gambolati, M. Putti and C. Paniconi, 'Projection methods for the finite element solution of the dual-porosity model in variably saturated porous media', in M.M. Aral (ed.), *Advances in Groundwater Pollution Control and Remediation*, vol. 9 of NATO ASI Series 2: Environment, Kluwer, Dordrecht, 1996, pp. 97–125.
10. H.A. van der Vorst, 'Bi-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of non-symmetric linear systems', *SIAM J. Sci. Stat. Comput.*, **13**, 631–644 (1992).
11. R.W. Freund, 'A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems', *SIAM J. Sci. Comput.*, **14**, 470–482 (1993).
12. D.S. Kershaw, 'The incomplete Cholesky-conjugate gradient method for the iterative solution of systems of linear equations', *J. Comp. Phys.*, **26**, 43–65 (1978).
13. Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, MA, 1996.
14. Y. Saad, 'ILUT: A dual threshold incomplete ILU factorization', *Numer. Lin. Alg. Appl.*, **1**, 387–402 (1994).
15. M. Benzi and M. Tuma, 'A comparative study of sparse approximate inverse preconditioners', *Tech. Rep. LA-UR98-0024*, Los Alamos National Laboratory, January, 1998.
16. M.T. van Genuchten and P.J. Wierenga, 'Mass transfer studies in sorbing porous media: 1. Analytical solutions', *Soil Sci. Soc. Amer. J.*, **40**, 473–480 (1976).
17. H.M. Leismann, B. Herrling and V. Krenn, 'A quick algorithm for the dead-end pore concept for modeling large-scale propagation processes in groundwater', in M.A. Celia, L. Ferrand, C.A. Brebbia, W.G. Gray and G.F. Pinder (eds.), *Proc. VII Int. Conf. on Computational Methods in Water Resources, vol. 2: Numerical Methods for Transport and Hydrologic Processes*, CMP Elsevier, Amsterdam, 1988, pp. 275–280.
18. E.A. Sudicky, 'The Laplace transform Galerkin technique: A time-continuous finite element theory and application to mass transport in groundwater', *Water Resour. Res.*, **25**, 1833–1846 (1989).
19. G. Galeati and G. Gambolati, 'On boundary conditions and point sources in the finite element integration of the transport equation', *Water Resour. Res.*, **25**, 847–856 (1989).
20. K.S. Crump, 'Numerical inversion of Laplace transforms using a Fourier series approximation', *J. Ass. Comput. Mach.*, **23**, 89–96 (1976).
21. F.R. de Hoog, J.H. Knight and A.N. Stokes, 'An improved method for numerical inversion of Laplace transforms', *SIAM J. Sci. Stat. Comput.*, **3**, 357–366 (1982).
22. C. Lanczos, 'Solution of systems of linear equations by minimized iterations', *J. Res. Natl. Bur. Stand.*, **49**, 33–53 (1952).
23. R. Fletcher, 'Conjugate gradient methods for indefinite systems', in *Lecture Notes Math.*, vol. 506, Springer, Berlin, 1976, pp. 73–89.
24. P. Sonneveld, 'CGS, a fast Lanczos-type solver for non-symmetric linear systems', *SIAM J. Sci. Stat. Comput.*, **10**, 36–52 (1989).
25. J.G.F. Francis, 'The QR transformation: A unitary analogue to the LR transformation', *Comp. J.*, **4**, 265–271 (1961).
26. E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov and D. Sorensen, *LAPACK User's Guide*, SIAM, Philadelphia, PA, 1992.

27. J.H. Wilkinson and C. Reinsch, *Handbook for Automatic Computation: Vol. II, Linear Algebra*, Springer, New York, 1971.
28. B.T. Smith, J.M. Boyle, J.J. Dongarra, V.S. Garbow, Y. Ikebe, V.C. Klema and C.B. Moler, 'Matrix eigensystem routines', in *Lecture Notes in Computer Science Vol. 6, EISPACK Guide*, Springer, New York, 1976.
29. F.L. Bauer, 'Der Verfahren der Treppeniteration und verwandte Verfahren zur losung algebraischer Eigenwert-probleme', *ZAMP*, 214–235 (1957).
30. Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, New York, 1992.
31. K.J. Bathe and E. Wilson, *Numerical Methods in Finite Element Analysis*, Prentice Hall, Englewood Cliffs, 1976.
32. H.C. Elman, 'A stability analysis of incomplete LU factorizations', *Math. Comp.*, **47**, 191–217 (1986).
33. G. Gambolati and G. Pini, 'Complex solution to non-ideal contaminant transport through porous media', *J. Comp. Phys.*, in press (1999).