Available online at www.sciencedirect.com
ScienceDirect
Applied Numerical Mathematics ••• (••••) •••–•••
APPLIED NUMERICAL MATHEMATICS
www.elsevier.com/locate/apnum

A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media

Burak Aksoylu [a], Hector Klie [b,*]

[a] *Department of Mathematics and Center for Computation and Technology, Louisiana State University, USA*
[b] *Center for Subsurface Modeling, Institute for Computational Science and Engineering, The University of Texas at Austin, USA*

Abstract

Eigenvalues of smallest magnitude become a major bottleneck for iterative solvers especially when the underlying physical properties have severe contrasts. These contrasts are commonly found in many applications such as composite materials, geological rock properties and thermal and electrical conductivity. The main objective of this work is to construct a method as algebraic as possible. However, the underlying physics is utilized to distinguish between high and low degrees of freedom which is central to the construction of the proposed preconditioner. Namely, we propose an algebraic way of separating binary-like systems according to a given threshold into high- and low-conductivity regimes of coefficient size $O(m)$ and $O(1)$, respectively where $m \gg 1$. So, the proposed preconditioner is essentially physics-based because without the utilization of underlying physics such an algebraic distinction, hence, the construction of the preconditioner would not be possible. The condition number of the linear system depends both on the mesh size Δx and the coefficient size m. For our purposes, we address only the m dependence since the condition number of the linear system is mainly governed by the high-conductivity sub-block. Thus, the proposed strategy is inspired by capturing the relevant physics governing the problem. Based on the algebraic construction, a two-stage preconditioning strategy is developed as follows: (1) a first stage that comprises approximation to the components of the solution associated to small eigenvalues and, (2) a second stage that deals with the remaining solution components with a deflation strategy (if ever needed). Due to its algebraic nature, the proposed approach can support a wide range of realistic geometries (e.g., layered and channelized media). Numerical examples show that the proposed class of physics-based preconditioners are more effective and robust compared to a class of Krylov-based deflation methods on highly heterogeneous media.
Published by Elsevier B.V. on behalf of IMACS.

Keywords: Preconditioning; Two-stage preconditioning; Schur complement; Physics-based preconditioning; Deflation; Heterogeneity; Porous media flow; Krylov subspace; GMRES; Multiscale; Iterative solver

1. Introduction

The main objective of the present work is to introduce a novel physics-based preconditioning strategy for solving diffusion-based problems in highly heterogeneous media. With the term physics-based we stress the fact that the physical connectivity described by the background media is exploited to define the algebraic character of our preconditioners. Hence, the proposed preconditioners benefit from both being still algebraic and using physical connectivity

* Corresponding author.
E-mail addresses: burak@cct.lsu.edu (B. Aksoylu), klie@ices.utexas.edu (H. Klie).

0168-9274/$30.00 Published by Elsevier B.V. on behalf of IMACS.
doi:10.1016/j.apnum.2008.06.002

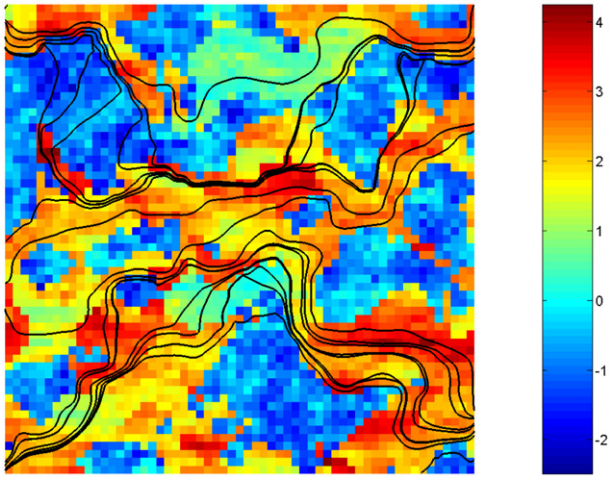

Fig. 1. Streamlines showing the preferential flow solution paths of the diffusion equation (1) on a channelized permeability field **K**.

information as an indicator to guide the solution process. These problems on highly heterogeneous media commonly arise, for example, in several material science, electrical conductivity and geosciences applications (see e.g., [23,26, 43,42,52]).

In recent years, understanding the role of systems characterized by highly heterogeneous media has been subject of intensive analysis via upscaling and multiscale procedures [16,26,40,58]. This fact has inspired new developments for solvers capable of capturing all possible scales present in the problem [2,22,21,24,30]. However, most of these developments are tightly dependent on the discretization method employed and none of them seem to particularly exploit representative or preferential patterns described by material properties and corresponding solution distributions. More precisely, these approaches disregard the fact that the random media often contain connected regions defining preferred solution paths (e.g., fluid flow, pressures, electrical conductance, resistivity or transmissibility depending on the application) that may be helpful to improve the solution process. Hence, there seems to be a gap in incorporating or retrieving most of the physical structure into the way linear systems are assembled and solved in highly variant diffusion problems.

Therefore, our proposed approach is strongly motivated by both the underlying physics and the simplicity than an algebraic treatment may offer. To that end, we are assuming that diffusion coefficients are characterized by a highly interconnected network that, in turn, yields a strong global conductivity media (e.g., channelized media or layered media joining the driving force of sink and source terms). Additionally, the media is assumed to be of almost binary character; that is, with distinguishable high- and low-conductivity regions. Fig. 1 illustrates the type of spatial conductivity distributions that we are interested in handling efficiently from the iterative solution standpoint. It represents a typical channelized permeability distribution on a porous media that generally, from moderate to large-scale problem sizes, imposes severe challenges to most modern solver technologies. A diffusion equation, depicted by Eq. (1), is used to reproduce the flow behavior on this channelized media with source and sink terms located at opposite extremes of the domain. Note that most of the relevant conductance or flow paths (indicated by the streamlines) is determined by the highly conductive network structure. The question that naturally arises is: Can we construct a simple algebraic strategy to account for these solution paths that may serve as reasonable approximation to the overall solution?

Thus, the first reasonable step in our approach is to consider that coefficients associated with distinguishable quasi-homogeneous conductivity regions should be grouped in the same block. In binary (or closely binary media) this means to separate matrix coefficients into high- and low- conductivity blocks thus giving rise to a 2×2 block linear system. The second step is to proceed with the solution of this linear system with a Krylov iterative solver (e.g., CG or GMRES) using a constrained solution to the high-conductivity block as a one preconditioning stage. Error components that still remain high after this stage can be additionally smoothed out by the action of a global preconditioning stage. This second stage should be cheap and is designed to capture solution components associated with the original coupling and possible roughness induced by the low-conductivity block. The rationale behind this two-stage precon-

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

ditioning strategy is that the high-conductivity block captures major response changes governing the solution of the overall coupled system. The procedure just described is purely algebraic and amenable for any discretization method.

Note that the procedure may be recursively applied in the event of multiple distinguishable regions, giving then rise to a multi-stage preconditioning method. Additionally, efficient multiscale solver approaches may be accommodated to take care of less varying heterogeneities contained in a particular coefficient block. In principle, the proposed method should be very appealing when the high-conductivity block is relatively small with respect to the size of the overall linear system.

Due to the target application considered here and the possible use of the deflation in any of the preconditioning stages, our development has parallel features to the pioneering deflation work of Vuik and several of his coauthors; see e.g., [18,38,55–57]. They particularly focus in the construction of deflation operators for layered problems with extreme permeability (conductivity) contrasts. One of their main contributions was to show that the set of smallest eigenvalues is in strong correspondence with the number of high-permeability layers surrounded by low-permeability layers. This result is a physical fact that aids at defining effective deflation strategies. Graham and Hagger [21, Section 5] gave rigorous analysis of the smallest eigenvalues arising when diagonal scaling is introduced. We intend to complement the existing work in the following two aspects: (1) provide a framework that may accommodate different deflation strategies as well as other preconditioning strategies such as coarse grid correction, domain decomposition and multigrid and, (2) allow for arbitrary variations within high- and low-conductivity regions. The proposed physics-based two-stage preconditioner is also inspired on previous work successfully applied for solving coupled linear systems involving different physical variables arising in fully implicit formulations of multiphase flow in porous media [10,15,27,51].

One can find an in-depth report of our approach that contains extensive numerical experiments in our preceding work [5]. The main emphasis of this work is to construct preconditioners that are robust with respect to the coefficient discontinuity, so we consider only moderate size linear systems. The robustness with respect to mesh size is the primary subject of our companion article [3]. In the case of linear finite element discretization, the article [3] contains rigorous analysis as well as extensions of our preconditioners to address mesh size robustness. More realistic applications of porous media flow and parallel computing implementation issues will be treated in the upcoming article [4].

The structure of the paper is as follows. Section 2 further emphasizes on the physical motivation supporting our approach. This includes a formalization of the matrix reorderings we are interested in constructing. Section 3 focuses in describing the two-stage preconditioning strategy designed to reflect the physics behind the linear systems. We show connections of these preconditioners with the Schur complement formulations and possible extensions to the proposed methodology. Also in Section 3, we show one of the main theoretical contributions. Namely, we establish that the m dependence of the preconditioned system is eliminated as $m \to \infty$. Section 4 relates Krylov-based and domain-based deflation approaches; including discussion on how to approximate eigenvectors and construct deflation operators for both approaches. This description has a two-fold blend: propose efficient ways to compute the second stage of two-stage preconditioners and revisit and compare some of the Krylov-based deflation operators that have been proposed in the literature. Section 5 numerically compares the proposed physics-based two-stage preconditioning strategy against some well established Krylov-based deflation methods. We end the paper with conclusions and directions of future work.

2. Problem formulation: exploiting the physics

2.1. Model problem

We consider the following simplified elliptic equation in our analysis:

$$-\nabla \cdot \big(\mathbf{K}(x)\nabla p\big) = q. \tag{1}$$

The conductivity coefficient $\mathbf{K}(x)$ depends on space and can be highly variant with possible discontinuities, but bounded below by a positive constant (i.e., coercivity condition). We also assume that Eq. (1) is defined on a domain Ω with non-flux boundary conditions, that is,

$$\nabla p \cdot \vec{n} = 0 \quad \text{on } \partial\Omega. \tag{2}$$

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

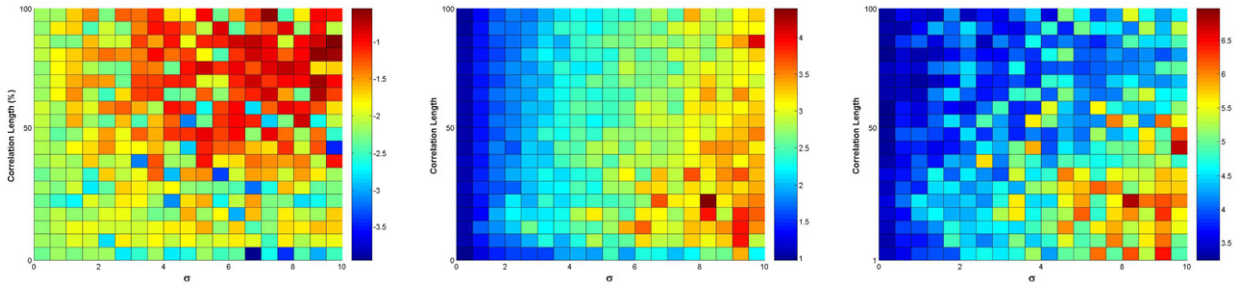

Fig. 2. Min eigenvalue, maximum eigenvalue and condition number as a function of variance and correlation length of $\mathbf{K}(x)$ according to an exponentially distributed random field.

Note that the solution is defined up to an arbitrary constant. Without loss of generality, invertibility of the resulting discretized system can be ensured by imposing $p = 0$ at any boundary gridblock. This will imply adding a positive constant to the matrix main diagonal. The solution of the equation is fundamentally driven by the sink/source term q. In many environmental and energy applications these terms are generally defined in regions where $\mathbf{K}(x)$ show higher values to facilitate the diffusion (or transport) process across the domain (e.g., increase sweep efficiency in cleanup strategies or oil recovery). Hence, $\mathbf{K}(x)$ plays a decisive role in the distribution of the solution p in the domain Ω. To simplify matters, we assume that $\mathbf{K}(x)$ is a positive scalar function describing highly conductive regions such as channels either branching or merging across Ω.

Fig. 2 shows the effect of heterogeneities in the value of extreme eigenvalues and in the resulting condition number of the linear system associated to Eq. (1). In this example, we consider the log of the diffusion coefficient $W(\mathbf{x}) = \ln K(\mathbf{x})$ following a second-order stationary distribution with a separable exponential covariance function of the form

$$C_W(\mathbf{x_1}, \mathbf{x_2}) = C_W(x_{1,1}, x_{1,2}; x_{2,1}, x_{2,2}) = \sigma_W^2 \exp\left[\frac{-1}{\eta}\{|x_{1,1} - x_{2,1}| + |x_{1,2} - x_{2,2}|\}\right], \tag{3}$$

where σ_W^2 is the spatial variability of the log field and η is the correlation length. Several realizations were generated for different combinations of σ_W^2 and η. Note that for a constant problem size and gridsize, the smaller the correlation length or the greater the spatial variability, the larger the matrix condition number. Similar trend can be obtained in other stationary and non-stationary distributions such as those arising in layered or channelized formations where the greater the coefficient jump (spatial variability) the higher the associated linear system condition number.

2.2. *Ordering unknowns according to permeability contrasts*

Since the coefficient $\mathbf{K}(x)$ defines highly conductive paths, it is convenient to partition Ω in two distinguishable regions. Let the open domains Ω_h and Ω_l denote the high-conductive and low-conductive regions, respectively. We assume that $\Omega_h \cup \Omega_l = \Omega$ and $\Omega_h \cap \Omega_l = \emptyset$ and let $\|\mathbf{K}(x)\|$ equal to $\mathrm{O}(m)$ and $\mathrm{O}(1)$, respectively in Ω_h and Ω_l where $m \gg 1$. Additionally, we assume that Ω_h is highly interconnected, that is, for any $x, y \in \Omega_h$ there is a continuous path joining these two points in the subdomain.

Hence, the following rationale defines the development of our forthcoming ideas: high-conductive regions should yield faster changes in the solution and conversely, low-conductive regions should yield almost constant solutions. Obviously, before putting this in practice we need to formalize what we mean by high- and low-conductive regions in a arbitrary binary or almost binary domain. This can be done by defining an average or threshold conductive value $\langle K \rangle$ that will be used to partition the domain in two distinguishable regions.

In general, the computation of $\langle K \rangle$ will depend on distribution of scales in Ω and it is a cornerstone issue for computing mean value fields on arbitrary media; see e.g., [42,50]. A practical choice for general random media is to compute $\langle K \rangle$ as the geometric mean, that is,

$$\langle K \rangle = \exp\left\{\frac{\int_\Omega \ln \mathbf{K}(x)\, d\mathbf{x}}{\int_\Omega d\mathbf{x}}\right\}. \tag{4}$$

The above expression is nothing more than a arithmetic mean in the log space. This definition of $\langle K \rangle$ is generally employed to find an effective value for a log-normal conductivity distribution. Nevertheless, we should remark that

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

geometric averaging is generally inaccurate to describe channelized media but sufficient for the type of cases we will analyze in this work. A more accurate computation of effective media in channelized media can be performed via percolative methods [45].

In this way, all grid elements with a conductivity value larger than $\langle K \rangle$ are numbered first and those with a value lower are numbered after. This gives rise to a 2×2 block system of the following form:

$$A^{\text{orig}} = \begin{bmatrix} A_h^{\text{orig}} & A_{hl}^{\text{orig}} \\ A_{lh}^{\text{orig}} & A_l^{\text{orig}} \end{bmatrix}. \tag{5}$$

In many practical situations A^{orig} is symmetric positive definite (SPD) and diagonal dominant. In fact, we will make this assumption throughout this work. A_h^{orig} and A_l^{orig} denote the blocks corresponding to high- and low-conductivity regions, respectively, A_{hl}^{orig} and A_{lh}^{orig} denote couplings. We have to emphasize on the use of the superscript "orig" to distinguish this system from the scaled one (to be described in the next section). Correspondingly, the Schur complement of A_h^{orig} in A^{orig} is given by $A_S^{\text{orig}} = A_l^{\text{orig}} - A_{lh}^{\text{orig}} A_h^{\text{orig}^{-1}} A_{hl}^{\text{orig}}$.

2.3. Properties and diagonal scaling

Given the block form (5), we expect to have concentrated the most relevant solution of the system in block A_h^{orig} since it captures most of the global conductivity of the system. To account for well defined conductivity values at each gridblock interface, it is necessary to average them accordingly. A distance-weighted harmonic averaging is usually performed to that end [46]. That is, for two neighboring blocks i and j, one approximates the directional conductivity K_{ij} at the interface ij as

$$K_{ij} = \frac{1}{2}(\Delta x_i + \Delta x_j)\left(\frac{\Delta x_i}{K_{ij,i}} + \frac{\Delta x_j}{K_{ij,j}}\right)^{-1}. \tag{6}$$

Applying a locally conservative discretization method such as control-volume [48] to (1) we obtain the following system of equations

$$\sum_j \frac{1}{2}\left(\frac{\Delta x_i}{K_{ij,i}} + \frac{\Delta x_j}{K_{ij,j}}\right)^{-1}(p_i - p_j) = \int_{\Omega_i} q\,dx, \tag{7}$$

for each element $\Omega_i \subset \Omega$. Note that, if we assume the discretization element size constant, the harmonic averaging imposes that A_h^{orig} is of order m whereas the other blocks are of order 1.

Some additional comments are in order:

- Given model equation (1), the coefficient blocks A_h^{orig} and A_l^{orig} are both symmetric positive definite (SPD) and diagonally dominant. This fact results from considering coercivity/boundedness assumptions and standard discretization procedures such as finite differences, finite elements or control volume.
- The dimension of A_h^{orig} and A_l^{orig} may differ significantly depending on the geometry. Important computational savings may be obtained when $|\Omega_l| \gg |\Omega_h|$.
- Each block A_h^{orig} and A_l^{orig} may consist of several main diagonal blocks associated with disconnected high-conductive regions (e.g., layering system where low- and high-conductive layers alternate in the formation). We should later stress that disconnected regions may not be necessarily included in the solution process.
- The number of layer blocks in each block A_h^{orig} and A_l^{orig} may vary in size and magnitude of its entries due to the presence of low-scale heterogeneities within each distinguishable conductivity zone.
- The blocks A_{hl}^{orig} and A_{lh}^{orig} tend to have smaller entries as the conductivity contrast becomes higher.

Since A^{orig} is SPD and diagonally dominant there are several implications in the properties associated with the blocks. First of all, note that these properties are invariant upon row and column permutations. Also, diagonal scaling creates a clustering effect and the spectral radius of the diagonally scaled matrix becomes $0 < \rho(A) < 2$ and difficulties in the solution process may arise when eigenvalues are either lying too close to the origin or 2.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

We invoke some important results associated with the 2×2 block partitioning given by (5) (see [6, Chapter 9], [25, p. 255]):

Theorem 2.1 *(On positive definiteness and condition number). If A^{orig} is SPD then each of the following statements follows*:

(1) A_h^{orig} *and* A_l^{orig} *are SPD.*
(2) A_S^{orig} *is SPD.*
(3) $\kappa_2(A_S^{\text{orig}}) \leqslant \kappa_2(A^{\text{orig}})$.
(4) $\| A_{lh}^{\text{orig}} A_h^{\text{orig}^{-1}} \|_2^2 \leqslant \kappa_2(A^{\text{orig}})$.

On the other hand, diagonal scaling brings up other important implications:

Lemma 2.1 *(On diagonal scaling). Let $D_l^{\text{orig}} = \text{diag}(A_l^{\text{orig}})$ and $D_h^{\text{orig}} = \text{diag}(A_h^{\text{orig}})$, then the following statements hold*:

(1) $S = D_l^{\text{orig}^{-1}} S^{\text{orig}}$.
(2) $\kappa_2(A_S) \leqslant \kappa_2(D_l^{\text{orig}}) \kappa_2(A_S^{\text{orig}})$.
(3) A_S *is similar to the SPD matrix* $T := D_l^{\text{orig}^{-1/2}} A_S^{\text{orig}} D_l^{\text{orig}^{-1/2}}$. *Therefore, eigenvalues of A_S are all positive.*
(4) *If A^{orig} is diagonally dominant, so is A.*

We denote the original system matrix by A^{orig}. Based on the just aforementioned facts, we construct a scaling or Jacobi preconditioner operator given by

$$D^{\text{orig}} = \begin{bmatrix} \text{diag}(A_h^{\text{orig}}) & 0 \\ 0 & \text{diag}(A_l^{\text{orig}}) \end{bmatrix} =: \begin{bmatrix} D_h^{\text{orig}} & 0 \\ 0 & D_l^{\text{orig}} \end{bmatrix}, \tag{8}$$

and compute the diagonally scaled system A:

$$A := D^{\text{orig}^{-1}} A^{\text{orig}} = \begin{bmatrix} D_h^{\text{orig}^{-1}} A_h^{\text{orig}} & D_h^{\text{orig}^{-1}} A_{hl}^{\text{orig}} \\ D_l^{\text{orig}^{-1}} A_{lh}^{\text{orig}} & D_l^{\text{orig}^{-1}} A_l^{\text{orig}} \end{bmatrix} = \begin{bmatrix} A_h & A_{hl} \\ A_{lh} & A_l \end{bmatrix}. \tag{9}$$

We observe that diagonal scaling improves the clustering of eigenvalues, therefore, diagonal scaling is always the initial default preconditioner. In various deflation preconditioner frameworks [56,57], the initial default preconditioner is chosen to be ILU(0). The simple diagonal scaling also conforms to our intention of keeping the preconditioner as simple as possible. Since diagonal scaling is the default preconditioner, all the block decomposition and analysis use subblocks from A not A^{orig}.

There is one more reason why we would like to use A instead of A^{orig}. Although A^{orig} is symmetric (hence normal), A is not necessarily normal. We would like to have a future extension of the two-stage preconditioner so that it works not only for symmetric matrices, but also for general non-normal matrices. To show the effectiveness of the two-stage preconditioner the test cases are simply limited to symmetric matrices.

3. Physics-based preconditioning

3.1. Two-stage preconditioning, the Schur complement, Δx and m dependence

Two stage-preconditioning refers to the concept of effectively combining the action of two preconditioners into a single one [14]. For instance, given the right-preconditioned matrix $A_1 = AM_1$, we can apply a second preconditioner (either from the right or left) to A_1. The idea can be easily generalized to multi-stage preconditioning by extending the recurrence several times. The motivation for combining multiple preconditioning stages obeys to the need of taking advantage of particular features of the problem within a single preconditioning stage. In a certain way, multigrid, and domain decomposition methods also fit into this framework. In the arena of porous media flow applications,

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

two-stage preconditioning has made a name by its own for the solution of fully coupled systems; see e.g. [15,27,51]. Moreover, this preconditioning technology is currently being considered an important initiative for the development of new generation of reservoir simulators for multiphase and compositional flow in porous media [10,19,33,54].

The two-stage preconditioner framework that we are interested in developing to be able to exploit the physics of the problem is the following:

Algorithm 3.1 *(Physics-based two-stage preconditioner).*

Solve high conductivity system: $A_h y_h = r_h$, where $A_h := R^t A R$, $r_h := R^t r$.
Obtain expanded solution: $y = R y_h$.
Compute new residual: $\hat{r} = r - Ay$.
Correct the residual: $\hat{v} = \hat{r} + y$.
(If needed) apply a stage two preconditioner M_d: $v = M_d^{-1}\hat{v}$.

The action of the whole preconditioner can be compactly written as

$$v = M_{\text{left}}^{-1} r = M_d^{-1}\big[I - (A - I)R(R^t A R)^{-1} R^t\big] r. \tag{10}$$

M_d^{-1} is an appropriate preconditioner of some desired kind such as a deflation preconditioner. In any case, the preconditioner M_d is used to solve those frequencies associated with the coupling to the low-conductivity block A_l. This step takes care of the small eigenvalues generated by the conductivity contrast at the interfaces. We note that M_{left} is an exact left inverse of A on the subspace spanned by the columns of R. That is, $(M_{\text{left}}^{-1} A)R = R$.

The inclusion operator under consideration is given as: $R = \left[\begin{smallmatrix} I_h \\ 0 \end{smallmatrix}\right]$. Then,

$$R(R^t A R)^{-1} R^t = \begin{bmatrix} A_h^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

If we do not use a stage two preconditioner, that is $M_d = I$, we obtain

$$M_{\text{left}}^{-1} = \begin{bmatrix} A_h^{-1} & 0 \\ -A_{lh} A_h^{-1} & I_l \end{bmatrix}. \tag{11}$$

Thus, M_{left}^{-1} becomes the exact inverse of $\left[\begin{smallmatrix} A_h & 0 \\ A_{lh} & I_l \end{smallmatrix}\right]$. If we further decompose (11) as

$$M_{\text{left}}^{-1} = \begin{bmatrix} A_h^{-1} & 0 \\ 0 & I_l \end{bmatrix} \begin{bmatrix} I_h & 0 \\ -A_{lh} A_h^{-1} & I_l \end{bmatrix}, \tag{12}$$

we then can connect M_{left}^{-1} to the following factorization of A,

$$A = \begin{bmatrix} I_h & 0 \\ A_{lh} A_h^{-1} & I_l \end{bmatrix} \begin{bmatrix} A_h & A_{hl} \\ 0 & A_S \end{bmatrix}, \tag{13}$$

with A_S denoting the Schur complement of A_h in A;

$$A_S = A_l - A_{lh} A_h^{-1} A_{hl}. \tag{14}$$

Now, combining (12) and (13) we obtain:

$$M_{\text{left}}^{-1} A = \begin{bmatrix} I_h & A_h^{-1} A_{hl} \\ 0 & A_S \end{bmatrix}, \tag{15}$$

indicating that

$$\sigma\big(M_{\text{left}}^{-1} A\big) = \sigma(A_S) \cup \{1\}. \tag{16}$$

Next, we will establish a condition number estimate in the Frobenius norm for the preconditioned system by utilizing the well-known identity:

$$\|G\|_F^2 = \text{trace}(G^t G). \tag{17}$$

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

We proceed to write $\|M_{\text{left}}^{-1}A\|_F^2$. For that, through (15) we easily see that

$$\left(M_{\text{left}}^{-1}A\right)^t M_{\text{left}}^{-1}A = \begin{bmatrix} I_h & A_h^{-1}A_{hl} \\ (A_h^{-1}A_{hl})^t & A_S^tA_S + (A_h^{-1}A_{hl})^t A_h^{-1}A_{hl} \end{bmatrix}.$$

Then,

$$\operatorname{trace}\left(\left(M_{\text{left}}^{-1}A\right)^t M_{\text{left}}^{-1}A\right) = \operatorname{trace}(I_h) + \operatorname{trace}\left(A_S^tA_S\right) + \operatorname{trace}\left(\left(A_h^{-1}A_{hl}\right)^t A_h^{-1}A_{hl}\right).$$

By (17), this is equivalent to the following:

$$\left\|M_{\text{left}}^{-1}A\right\|_F^2 = N_h + \|A_S\|_F^2 + \left\|A_h^{-1}A_{hl}\right\|_F^2, \tag{18}$$

where N_h denotes the number of degrees of freedom (DOF) associated to $\bar{\Omega}_h$.

Then, we start writing an expression for $\|(M_{\text{left}}^{-1}A)^{-1}\|_F^2$. First, (15) implies that

$$\left(M_{\text{left}}^{-1}A\right)^{-1} = \begin{bmatrix} I_h & -A_h^{-1}A_{hl}A_S^{-1} \\ 0 & A_S^{-1} \end{bmatrix}. \tag{19}$$

Using (19), we get

$$\left(M_{\text{left}}^{-1}A\right)^{-t}\left(M_{\text{left}}^{-1}A\right)^{-1} = \begin{bmatrix} I_h & -A_h^{-1}A_{hl}A_S^{-1} \\ (-A_h^{-1}A_{hl}A_S^{-1})^t & A_S^{-t}A_S^{-1} + (A_h^{-1}A_{hl}A_S^{-1})^t A_h^{-1}A_{hl}A_S^{-1} \end{bmatrix}.$$

Using a similar argument as in (18), we reach:

$$\left\|\left(M_{\text{left}}^{-1}A\right)^{-1}\right\|_F^2 = N_h + \left\|A_S^{-1}\right\|_F^2 + \left\|A_h^{-1}A_{hl}A_S^{-1}\right\|_F^2. \tag{20}$$

Using (18) and (20), the condition number follows:

$$\operatorname{cond}_F\left(M_{\text{left}}^{-1}A\right) = \left\{N_h + \|A_S\|_F^2 + \left\|A_h^{-1}A_{hl}\right\|_F^2\right\}^{1/2}\left\{N_h + \left\|A_S^{-1}\right\|_F^2 + \left\|A_h^{-1}A_{hl}A_S^{-1}\right\|_F^2\right\}^{1/2}. \tag{21}$$

In the general setting, cond_F is a function of both the mesh size Δx and the coefficient size m. The dependence on Δx is investigated in the companion article [3]. Therefore, in this article we always assume that Δx is fixed and address only the m dependence.

We would like to report the following results (24) and (25) from [3] for linear finite element discretization of (1) with purely homogeneous Dirichlet boundary condition (i.e. $p = 0$ on $\partial\Omega$) in which Ω_h is an isolated island; $\bar{\Omega}_h \cap \partial\Omega = \emptyset$. This setting is the simplest possible representation of the highly heterogeneous media that encapsulates the effects of a high-conductive region. Incorporating the m dependence to all the relevant matrices, we see that the block representation in (5) is in fact as follows:

$$A^{\text{orig}}(m) = \begin{bmatrix} A_h^{\text{orig}}(m) & A_{hl}^{\text{orig}} \\ A_{lh}^{\text{orig}} & A_l^{\text{orig}} \end{bmatrix}. \tag{22}$$

This decomposition assumes that high-conductive subblock captures all the m dependent DOF and the rest of the subblocks are m independent. Then, the condition number estimate in (21) easily extends to the case without diagonal scaling:

$$\begin{aligned}\operatorname{cond}_F\left(M_{\text{left}}^{\text{orig}^{-1}}(m)A^{\text{orig}}(m)\right) = {} & \left\{N_h + \left\|A_S^{\text{orig}}(m)\right\|_F^2 + \left\|A_h^{\text{orig}^{-1}}(m)A_{hl}^{\text{orig}}\right\|_F^2\right\}^{1/2} \\ & \times \left\{N_h + \left\|A_S^{\text{orig}^{-1}}(m)\right\|_F^2 + \left\|A_h^{\text{orig}^{-1}}(m)A_{hl}^{\text{orig}}A_S^{\text{orig}^{-1}}(m)\right\|_F^2\right\}^{1/2}.\end{aligned} \tag{23}$$

Next, we show that asymptotically m dependence is eliminated.

Theorem 3.1.

$$\lim_{m\to\infty} A_h^{\text{orig}^{-1}}(m) = A_{h,\infty}^{\text{orig}^*}, \tag{24}$$

$$\lim_{m\to\infty} A_S^{\text{orig}}(m) = A_{S,\infty}^{\text{orig}}. \tag{25}$$

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

Remark 3.1. In our recent work [3], in the case of linear finite element discretization, we completely reveal the above limiting matrices. In particular, $A_{h,\infty}^{\text{orig}^*}$ and $A_{S,\infty}^{\text{orig}}$ converge to a low rank matrix and a low rank perturbation of A_l^{orig}, respectively, with the same rank. The rank is determined by the number of disconnected components (islands) forming the high permeable region. Furthermore, we can explicitly determine the rate of convergence. For sufficiently large m, we have:

$$A_h^{\text{orig}^{-1}}(m) = A_{h,\infty}^{\text{orig}^*} + \mathcal{O}(m^{-1}),$$
$$A_S^{\text{orig}}(m) = A_{S,\infty}^{\text{orig}} + \mathcal{O}(m^{-1}).$$

Utilizing the fact that both A^{orig} and $A^{\text{orig}^{-1}}$ are continuous in m [21, Lemma 5.1(i)] together with the continuity of $\|\cdot\|_F$, the immediate Corollary shows that asymptotically m dependence is eliminated for the condition number as well.

Corollary 3.1.

$$\lim_{m\to\infty} \text{cond}_F\big(M_{\text{left}}^{\text{orig}^{-1}}(m)A^{\text{orig}}(m)\big) = \big\{N_h + \big\|A_{S,\infty}^{\text{orig}}\big\|_F^2 + \big\|A_{h,\infty}^{\text{orig}^*}A_{hl}^{\text{orig}}\big\|_F^2\big\}^{1/2} \times \big\{N_h + \big\|A_{S,\infty}^{\text{orig}^{-1}}\big\|_F^2 + \big\|A_{h,\infty}^{\text{orig}^*}A_{hl}^{\text{orig}}A_{S,\infty}^{\text{orig}^{-1}}\big\|_F^2\big\}^{1/2}. \tag{26}$$

Asymptotically, we have showed that the condition number of the preconditioned system becomes independent of m. In other words, the preconditioner eliminates the m dependence in the limiting case. Using the identity,

$$A_S(m) = D_l^{\text{orig}^{-1}} A_S^{\text{orig}}(m),$$

and the m independence of $D_l^{\text{orig}^{-1}}$, a condition number estimate similar to the one in (26) can be obtained for $M_{\text{left}}^{-1}(m)A(m)$. We have used a finite volume discretization for our numerical results. The above results for linear finite element discretization easily carry over to cell centered finite volume discretization as well as to cell centered finite difference because all three are equivalent when the PDE has constant coefficients.

If we supplement M_{left}^{-1} in (11) by including the inversion of A_S, then an "ideal" preconditioner can be obtained that here it is defined as $M_{\text{left,opt}}^{-1}$:

$$M_{\text{left,opt}}^{-1} = \begin{bmatrix} A_h^{-1} & 0 \\ 0 & A_S^{-1} \end{bmatrix} \begin{bmatrix} I_h & 0 \\ -A_{lh}A_h^{-1} & I_l \end{bmatrix}, \tag{27}$$

$$M_{\text{left,opt}}^{-1}A = \begin{bmatrix} I_h & A_h^{-1}A_{hl} \\ 0 & I_l \end{bmatrix}, \tag{28}$$

with $\sigma(M_{\text{left,opt}}^{-1}A) = \{1\}$. The availability of an effective preconditioner of A_S in our two-stage preconditioning is highly desirable. This question has been heavily studied. We note that A_S loses sparsity due to the inversion of A_h. Moreover, A_S can be ill-conditioned with respect to Δx. The ultimate performance of our two-stage preconditioner hinges on the effectiveness of the preconditioner of A_S. But the favorable feature of our preconditioner is that asymptotically m dependence is eliminated, we only ask for an effective preconditioner for A_S with respect to Δx only. We conclude that the dependence of the Schur complement is the inherit property of our two-stage preconditioner.

An important aspect is that the formulation of (10) may be applied in a nested fashion. That is, we can repeat the above procedure for the solution of the A_h. This may be useful if the high-conductivity block may still have strong variations within. This implies to solve a hierarchy of subproblems for which we calculate $\langle K\rangle_i$ at each level i and apply the deflation M_d step as a smoothing step. The procedure is repeated until the underlying subproblem either resembles a purely homogeneous case or is sufficiently small to enable the use of a direct or fast iterative solver.

The preconditioner defined above belongs to an extensive family of two-stage preconditioners [10,15,27,51] for solving fully-coupled linear systems where each of the variables involved follow different physical behavior (e.g., pressures and saturations sharing the same discretization block in multiphase flow scenario). We have been inspired by this idea to define the above preconditioner strategy to simultaneously accommodate full high-conductivity solutions

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

with deflated low-conductivity solutions. To our knowledge, this is the first time that a two-stage preconditioner approach has been specifically used to deal with high conductivity contrasts as it may occur in (close) binary media.

3.2. Left versus right preconditioner

Let us consider the below decompositions of A:

$$A = \left\{ \begin{bmatrix} I_h & 0 \\ A_{lh}A_h^{-1} & I_l \end{bmatrix} \begin{bmatrix} A_h & 0 \\ 0 & I_l \end{bmatrix} \right\} \begin{bmatrix} I_h & 0 \\ 0 & A_S \end{bmatrix} \begin{bmatrix} I_h & A_h^{-1}A_{hl} \\ 0 & I_l \end{bmatrix}$$
$$= M_{\text{left}} \begin{bmatrix} I_h & A_h^{-1}A_{hl} \\ 0 & A_S \end{bmatrix}, \tag{29}$$
$$A = \begin{bmatrix} I_h & 0 \\ A_{lh}A_h^{-1} & I_l \end{bmatrix} \begin{bmatrix} I_h & 0 \\ 0 & A_S \end{bmatrix} \left\{ \begin{bmatrix} A_h & 0 \\ 0 & I_l \end{bmatrix} \begin{bmatrix} I_h & A_h^{-1}A_{hl} \\ 0 & I_l \end{bmatrix} \right\}$$
$$= \begin{bmatrix} I_h & 0 \\ A_{lh}A_h^{-1} & A_S \end{bmatrix} M_{\text{right}}. \tag{30}$$

We can define a left and right preconditioner from (29) and (30), respectively:

$$M_{\text{left}}^{-1} = \begin{bmatrix} A_h^{-1} & 0 \\ -A_{lh}A_h^{-1} & I_l \end{bmatrix}, \tag{31}$$
$$M_{\text{right}}^{-1} = \begin{bmatrix} A_h^{-1} & -A_h^{-1}A_{hl} \\ 0 & I_l \end{bmatrix}. \tag{32}$$

Then,

$$M_{\text{left}}^{-1}A = \begin{bmatrix} I_h & A_h^{-1}A_{hl} \\ 0 & A_S \end{bmatrix}, \tag{33}$$
$$AM_{\text{right}}^{-1} = \begin{bmatrix} I_h & 0 \\ A_{lh}A_h^{-1} & A_S \end{bmatrix}. \tag{34}$$

We conclude that the spectra of the preconditioned systems (33) and (34) are the same:

$$\sigma\left(M_{\text{left}}^{-1}A\right) = \sigma\left(AM_{\text{right}}^{-1}\right) = \sigma(A_S) \cup \{1\}. \tag{35}$$

Moreover, actions of both (31) and (32) require only one action of A_h^{-1}.

$$M_{\text{left}}^{-1} \begin{bmatrix} x_h \\ x_l \end{bmatrix} = \begin{bmatrix} (A_h^{-1}x_h) \\ x_l - A_{lh}(A_h^{-1}x_h) \end{bmatrix},$$
$$M_{\text{right}}^{-1} \begin{bmatrix} x_h \\ x_l \end{bmatrix} = \begin{bmatrix} A_h^{-1}(x_h + A_{hl}x_l) \\ x_l \end{bmatrix}.$$

If a non-trivial stage two preconditioner is introduced, then the preconditioned systems in (33) and (34) will take the forms

$$M_d^{-1}\left(M_{\text{left}}^{-1}A\right), \tag{36}$$
$$\left(AM_{\text{right}}^{-1}\right)M_d^{-1}. \tag{37}$$

In order to have a spectrally equivalent preconditioned system as in (35), the left preconditioner should be in the immediate left side of A. Similarly, the right preconditioner should be in the immediate right side of A. In (36), M_{left}^{-1} and M_d^{-1} become the stage one and stage two preconditioners, respectively. The order of preconditioners is changed in the right preconditioned case. In (37), M_d^{-1} and M_{right}^{-1} become the stage one and stage two preconditioners, respectively. We note that the order of stage one and stage two preconditioners determines the main difference between left and right preconditioned cases.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

3.3. Spectral analysis and matrix conditioning

Since A^{orig} is symmetric positive definite (SPD), all the related subblocks become SPD. Hence, the condition numbers are simply the ratio of the maximum eigenvalue over the minimum one. However, when diagonal scaling is introduced, $A := D^{\text{orig}^{-1}} A^{\text{orig}}$, A is not symmetric, hence, $\kappa(A)$ is not necessarily equal to $\lambda_{\max}(A)/\lambda_{\min}(A)$. A natural question arises: How are $\kappa(A)$ and $\lambda_{\max}(A)/\lambda_{\min}(A)$ related?

Since A^{orig} is SPD, then $D^{\text{orig}^{1/2}}$ is defined as a real matrix. In order to make this connection, let us introduce $A^{\text{sym}} := D^{\text{orig}^{-1/2}} A^{\text{orig}} D^{\text{orig}^{-1/2}}$. Note that A and A^{sym} are similar through

$$A = D^{\text{orig}^{-1/2}} A^{\text{sym}} D^{\text{orig}^{1/2}}. \tag{38}$$

Then, they share the same spectrum:

$$\sigma(A) = \sigma(A^{\text{sym}}). \tag{39}$$

Using (39) and the fact that A^{sym} is SPD, we get the following:

$$\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \frac{\lambda_{\max}(A^{\text{sym}})}{\lambda_{\min}(A^{\text{sym}})} = \kappa(A^{\text{sym}}). \tag{40}$$

Since A and A^{sym} are similar, using (38) and the fact that $\kappa^2(D^{\text{orig}^{1/2}}) = \kappa(D^{\text{orig}})$, we get an upper and a lower bound for $\kappa(A)$

$$\frac{1}{\kappa(D^{\text{orig}})}\,\kappa(A^{\text{sym}}) \leqslant \kappa(A) \leqslant \kappa(D^{\text{orig}})\,\kappa(A^{\text{sym}}). \tag{41}$$

Eq. (40) together with (41) reveal the connection we are after:

$$\frac{1}{\kappa(D^{\text{orig}})}\,\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leqslant \kappa(A) \leqslant \kappa(D^{\text{orig}})\,\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}. \tag{42}$$

$\lambda_{\max}(A)/\lambda_{\min}(A)$ provides a good estimate of $\kappa(A)$ when $\kappa(D^{\text{orig}}) \approx 1$. Since our conductivity field is assumed to be highly heterogeneous, this will seldom happen. Therefore, in general $\lambda_{\max}(A)/\lambda_{\min}(A)$ does not reflect $\kappa(A)$. Furthermore, subblocks of A are related to that of A^{orig} in a similar fashion.

In short, the spectral values of A and its subblocks do not reveal much in terms of the condition number. That is why we directly computed condition number values as given in Table 1. However, it is not just the condition number that plays the essential role for the convergence of solvers. The distribution of the eigenvalues is crucial as well. Especially, smallest eigenvalues in spectral plots will help us reveal certain aspects of the solver performance.

Due to (16), we would like to emphasize that the spectrum of A_S directly dictates the spectrum of the preconditioned system. The spectral plot of A_S provide indications to justify the use of a stage two preconditioner. Namely, if one observes an outlier smallest eigenvalue in spectral plot of A_S, this is the exactly the same eigenvalue which could not be resolved by the stage one preconditioner. Then we can consider to use a stage two preconditioner to address the complications that the outlier smallest eigenvalues can cause. We offer to exploit deflation methods as they are known to be effective to attenuate the complications of outlier eigenvalues may cause. Such outlier eigenvalues arise especially when the A_h block misses DOF that are associated to the high-conductive region; see Figs. 4 and 6.

4. Deflation methods

4.1. Fundamentals

In the recent years, deflation methods have been increasingly receiving attention as a way for improving the convergence of linear iterative solvers. Deflation operators provide means to remove the negative effect that extreme (usually small) eigenvalues have in the convergence of Krylov iterative methods for solving symmetric and non-symmetric systems [9,13,17,20,35–37,47,49]. In most of these research efforts, deflation methods have been developed as a mechanism for systematically expanding and refreshing the underlying Krylov subspace or for conceiving more effective preconditioning techniques. Nevertheless, the use of deflation is not strictly confined to the setting of Krylov subspace iterations (e.g., [8,18,57,24]) or even to the solution of linear system of equations (e.g., [7,28,41]).

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

In addition, a given preconditioner can be complemented with a deflation procedure. In fact, deflation can be used to take care of the "roughness" that was left out upon, for instance, decoupling the system. Thus, deflation allows for knocking down components of the solutions associated with both small and large eigenvalues generated by the conductivity contrast. The two well-known strategies for deflation are the Krylov-based and the domain-based methods. The former utilizes eigenspace information (in the form of Ritz and harmonic Ritz vectors) from the underlying Krylov iterative solver to construct the deflating subspace; see e.g. [20,36]. We can also identify this Krylov-based method as dynamic deflation methods as they accuracy can vary depending on the closeness to an invariant subspace that the approximating Krylov subspace entails. In this case, the deflation operator is regularly updated as the fresh Krylov subspace information is computed.

Domain-based approaches exploit coefficient contrasts to identify subdomain blocks in which the solution demonstrates a certain behavior. Using the two information together geometric or algebraic operators are constructed to approximate eigenvectors; see [18,55]. This makes the deflation highly dependent on the domain and the behavior of the solution. The main advantage of domain-based deflation methods is that they do not rely on the explicit computation of approximate eigenvalues and eigenvectors. In this sense, domain-based deflation can be also identified as static deflation since the deflation operator is determined before the iteration process starts and remains fixed throughout.

A typical deflation operator is designed to process the extremal eigenvalues in such a way that the resulting operator will a have better condition number in overall. This goal can be accomplished in many ways. For instance, mapping the smallest eigenvalues to 0 or shifting them to 1 or $\lambda_{\max}(A)$. Let $U \in \mathbb{C}^{n\times r}$ be the exact invariant subspace corresponding to r smallest eigenvalues.

One type of deflation operator that shifts the r smallest eigenvalues to $|\lambda_{\max}(A)|$ and leaves the rest of the spectrum unchanged is given by [8]:

$$C^{-1} = \big|\lambda_{\max}(A)\big|U(U^TAU)^{-1}U^T + (I - UU^T), \tag{43}$$

where $|\lambda_{\max}(A)|$ is the magnitude of the largest eigenvalue. We utilize the operator in (43) as the stage two preconditioner. If U is a near invariant subspace, then depending on the approximation quality of U, the preconditioned matrix AC^{-1} will have eigenvalues close to the set $\{\lambda_{r+1}, \ldots, \lambda_{\max}(A), |\lambda_{\max}(A)|, \ldots, |\lambda_{\max}(A)|\}$.

A preconditioning technique which aims at utilizing the spectral information when restarting would be ideal. The idea is to compute a near invariant subspace corresponding to the smallest eigenvalues. Indeed, the rate of convergence is mostly governed by these smallest eigenvalues [8,17]. The full-GMRES version behaves as if the smallest eigenvalues were removed after some iterations which is typically characterized by superlinear converge. But this is no longer true in the restarted case. Therefore, we remove them with the help of a deflation preconditioner. After each cycle of GMRES(m), the preconditioner is updated by pulling out new eigenvalues. At each restart, new approximate eigenvectors are estimated in order to increase the quality of the invariant subspace.

In spite of all the aforementioned advances, the idea of deflating unwanted eigenvectors from the solution is not new. During the end of the 60's, all 70's and beginning of the 80's, deflation was primarily employed for constructing meaningful solutions for (almost) singular linear systems [11,12,29,44,53]. A novel view of the approach was provided by Nicolaides [39] who fundamentally propose to split the conjugate gradient solution as the sum of a deflated subspace conjugate solution plus a particular solution into the complementary subspace. Nicolaides realized that the new procedure was amenable to use in conjunction with other preconditioners techniques. The idea was further explored by Mansfield in the setting of domain decomposition [31,32].

4.2. Deflation methods under consideration

We consider three well-known Krylov-based deflation methods: (1) harmonic [17], (2) augmented [34], and (3) Burrage and Erhel [8]. In the numerical experiments, we will denote these methods as `harmo, aug, BE`, respectively. All the methods in this article are implemented as preconditioners to `gmres(m)`.

All the deflation methods under consideration utilize a near invariant subspace extracted from the Hessenberg matrix produced by the underlying `gmres(m)` method. Chapman and Saad [13] report that harmonic projection produces approximate eigenvectors that are more accurate than the ones produces by an oblique projection. We prefer to compute the near invariant subspace corresponding to smallest eigenvalues by *harmonic Ritz projection* due to its favorable approximation properties of the eigenvectors corresponding to smallest eigenvalues.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

In Algorithms 4.1 and 4.2, we provide the description of the harmonic and augmented deflation methods in the same spirit on which they were presented in [8]. The Burrage–Erhel deflation was introduced in [17] and it is an improved version of the harmonic deflation method. Algorithmically, it is identical to harmonic deflation except that the harmonic Ritz projection is used in the following improved way. A near invariant subspace U_{old} is computed by using harmonic projection of the eigen-residual onto the Krylov subspace just like it is done in the harmonic deflation method. At the end of the cycle, U_{old} is retained and a fresh near invariant subspaces, U_{fresh}, is computed, then orthogonalized against U_{old}. They are appended to form a bigger subspace $[U_{\text{old}}, U_{\text{fresh}}]$. At the end, an other harmonic projection of the eigen-residual is performed onto $[U_{\text{old}}, U_{\text{fresh}}]$ to form the new invariant subspace U_{new}. This way of updating U is slightly more costly but more dynamic and seems to give better convergence rates.

Algorithm 4.1 *(***harmo**(m, l) *(Deflation by eigenvalue shift)).*

convergence := false;
choose x_0;
$C := I_N$;
$U := [\,]$;
until convergence do
  $r_0 = b - Ax_0$;
  apply Arnoldi process to AC^{-1} to compute V_m;
  $y_m = \text{argmin}_{y \in \Re^m} \|\beta e_1 - \bar{H}_m y\|$;
  $x_m := x_0 + C^{-1} V_m y_m$;
  if $\|b - Ax_m\| <$ tolerance; convergence := true;
  else
    $x_0 = x_m$;
    compute an invariant subspace U of A of size l by harmonic projection;
    compute $C^{-1} = |\lambda_{\max}(A)| U (U^T A U)^{-1} U^T + (I - UU^T)$;
  endif;
enddo;

In the original construction the harmonic and Burrage–Erhel methods are designed to be right preconditioners. In the right preconditioned case, the system becomes $AC^{-1}u = b$, where $x = C^{-1}u$. The residual expression is the following:

$$r_i = b - AC^{-1}u_i = b - Ay_i.$$

Assuming a restart at every m-th iteration, at the i-th iteration the GMRES algorithm finds $\|\bar{r}_i\| = \|b - Ax_i\|$:

$$\|\bar{r}_i\| = \min_{y_i \in \kappa_{\mathcal{J}(i)}(A, r_0)} \|r_i\|,$$

where

$$\mathcal{J}(i) = \begin{cases} i, & i \neq 0 \bmod m, \\ m, & i = 0 \bmod m \end{cases}$$

and in the next iteration $i + 1$, the vector $A^{\mathcal{J}(i+1)-1} r_0$ is added to Krylov subspace. Within each cycle, the Krylov subspace over which we search the minimum gets larger, this implies the well-known non-increasing residual property:

$$\|\bar{r}_{i+1}\| \leqslant \|\bar{r}_i\|, \quad \text{for } i \text{ that stays in the same cycle.} \tag{44}$$

After a restart, for instance at iteration $m + 1$, computations take place in then new cycle as we look for

$$\|\bar{r}_{m+1}\| = \min_{y_{m+1} \in \kappa_1(A, r'_0)} \|r_{m+1}\|, \tag{45}$$

where $r'_0 = \bar{r}_m$ due to assigning $y_{m+1} := x_m$ before the start of the new cycle. (45) implies that $\|\bar{r}_{m+1}\| \leqslant \|b - Ay_{m+1}\|$ for any $y_{m+1} \in \kappa_1(A, r'_0)$, in particular for x_m. Therefore, in exact arithmetic, GMRES(m) generates a non-increasing residual for the iteration between two cycles as well:

$$\|\bar{r}_{m+1}\| \leqslant \|b - Ax_m\| = \|\bar{r}_m\|. \tag{46}$$

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

Algorithm 4.2 *(***aug**(m, l) *(Deflation by augmenting)).*

convergence := false;
choose x_0;
$U := [\,]$;
until convergence do
  $r_0 = b - Ax_0$;
  apply Arnoldi process to A to compute V_m;
  $W = [V_m, U]$;
  compute AU;
  orthogonalize AW to get V;
  $y_m = \text{argmin}_{y \in \Re^{m+l}} \|\beta e_1 - \bar{H} y\|$;
  $x_m := x_0 + W y_m$;
  if $\|b - Ax_m\| <$ tolerance; convergence := true;
  else
    $x_0 = x_m$;
    compute an invariant subspace U of A of size l by harmonic projection;
  endif;
enddo;

In Krylov-based or dynamic deflation, the preconditioner C^{-1} is updated at every cycle. But the residual expression does not contain C^{-1} in the right preconditioned case. In the left preconditioner case, the system is $C^{-1}Ax = C^{-1}b$. Then, $r_i = C^{-1}b - C^{-1}Ay_i = b - Ay_i$. In the deflation methods used, C^{-1} depends on the near invariant subspace. Since, the near invariant subspace is updated at the end of each cycle, the preconditioner does not change within the cycle. At the restart, C^{-1} is updated, hence (46) does not hold. So, in exact arithmetic, one can expect increase at the restart but (44) will hold within each cycle. This explains the oscillations we observe for the left preconditioned case.

5. Numerical experiments

We have chosen a set of 4 different numerical experiments to establish that our physics-based two-stage preconditioner is an effective alternative for randomly heterogeneous applications. Since the preconditioner is a two-stage preconditioner it has two components. The main component (in the left preconditioned case) is the preconditioner in (11) or in (31). If needed, a stage two preconditioner can be employed as in (36) to complement the stage one preconditioner. The stage two preconditioner chosen is the harmonic method. Our two-stage preconditioner has also a version based on right preconditioning. The one stage right preconditioner is given in (32) and its two stage counterpart is given (37). The right preconditioner will give rise to non-increasing residual in exact arithmetic whereas the residual in the left preconditioner can have oscillations because the residual at the restarts can increase due to the updated near invariant subspace as explained in Section 4.2.

The methods employed in this paper have been implemented in MATLAB 7.1. We define the preconditioner effectiveness as the rate of decay of the relative residual with respect to matrix-vector multiplications (MVPs). For each test problem, we report the residual and error convergence history and the iteration counts for the left and the right preconditioned cases. In a given iteration plot, we report only the methods that converged. If there is no corresponding iteration count, this means that the method did not meet the convergence criterion. The stopping criterion is chosen to be $\|\frac{r_m}{r_0}\|_2 < 1.0\text{e} - 10$.

Table 1 summarizes the characteristics associated with each of the 4 cases considered. As we can see, the cases represent different conductivity distribution geometries and contrasts. Consequently, the condition number of the associated matrix varies relatively high despite the modest size of the cases considered. We have listed the condition numbers associated with the most relevant blocks without and with diagonal scaling. We can clearly observe the crucial role that scaling and the proposed physics-based two-stage preconditioning play in decreasing the condition number of the original matrix. In the results shown in the table, $M_d = I$ (i.e., the two-stage preconditioner only includes high-conductivity block solutions). We employed a finite-volume discretization method [48]. Hence, all of the system matrices, A^{orig}, are symmetric positive definite, diagonally dominant, and highly ill-conditioned.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

Table 1
The numbers of degrees of freedom and condition numbers

Test problem	1	2	3	4
N	1600	1600	1600	1600
N_h; N_l	340; 1260	457; 1143	462; 1138	562; 1038
$K_{\min}$	1.00e − 04	1.00e − 07	1.00e − 02	3.16e − 02
$K_{\max}$	1.00e + 04	3.98e + 04	1.00e + 05	1.00e + 05
$\langle K \rangle$	2.20e − 03	2.29e − 02	1.66e − 01	6.03e + 00
cond(A)	1.65e + 05	1.76e + 13	1.33e + 10	5.89e + 10
cond(A_h)	2.92e + 04	4.52e + 10	4.58e + 08	8.43e + 06
cond(A_l)	1.64e + 03	1.05e + 03	1.91e + 04	8.08e + 00
cond(A_S)	1.77e + 03	3.17e + 06	3.62e + 04	1.34e + 04
cond($M_{\text{left}}^{-1}A$)	2.00e + 03	4.69e + 07	3.18e + 05	3.23e + 04
cond(A^{orig})	1.65e + 11	8.29e + 12	1.57e + 10	3.20e + 10
cond(A_h^{orig})	4.35e + 08	1.61e + 12	2.62e + 08	7.37e + 06
cond(A_l^{orig})	1.52e + 03	3.22e + 05	2.08e + 04	1.07e + 01
cond(A_S^{orig})	1.64e + 03	1.89e + 06	3.56e + 04	1.09e + 04
cond($M_{\text{left}}^{\text{orig}^{-1}}A^{\text{orig}}$)	2.07e + 06	1.60e + 09	4.58e + 05	6.90e + 04

5.1. Spectral analysis and effects of diagonal scaling

In this section, we outline some of the connections between the conductivity field and the spectrum. There are only two conductivity values (high and low value) in test problems 1 and 4, whereas in test problems 2 and 3 we allow variation in the conductivity values within the high and low conductivity regions to analyze the reliability scope of the proposed approach. There is several orders of magnitude difference between these eigenvalues similar to the conductivity values. The entries of A^{orig} corresponding to $A_h^{\text{orig}}(m)$ are O(m) and that of A_l^{orig} are O(1) due to high values of conductivity.

The severe contrast in conductivity values creates two main clusters of eigenvalues of A^{orig}: large and small. Our immediate observation is that the large and small eigenvalues of A^{orig} correspond to eigenvalues of A_h^{orig} and A_l^{orig}, respectively. The contrast becomes more apparent when the conductivity values are homogeneous within the high and low conductive regions (see Figs. 3, 6) and the separation diminishes with less homogeneous conductivity values (see Fig. 5) and almost disappears (see Fig. 4). Comparing Figs. 3, 4, and 5, we observe that the number of shared eigenvalues increase as the variation in the conductivity values increases.

When diagonal scaling is introduced, eigenvalues of A_h capture most of the main features of the spectrum of A, especially the smallest eigenvalues of A; see e.g., Fig. 3. We observe that diagonal scaling greatly helps collecting smallest eigenvalues of A in A_h. Therefore, when we introduce diagonal scaling, the main motivation is to find smallest eigenvalues of A that are responsible for conductivity contrasts. Once these small eigenvalues are found, the preconditioner will target them to eliminate their complications for the solver. The existing works by Vuik et al. [56,57] and Graham and Hagger [21] explain the effects of diagonal scaling in the case of diffusion equation with highly contrasting diffusion coefficients. In the particular case of layered media, conductivity regions that are sandwiched by low conductivity region provides the exact number of smallest eigenvalues of A.

Next, we report some of our observations and state same heuristics on the behavior of smallest eigenvalues in the spectrum. The number of well-identifiable high conductivity regions surrounded by low conductivity regions corresponds to the number of smallest eigenvalues of A. In all these cases, smallest eigenvalues of A are captured as the smallest eigenvalues of A_h subblock. For instance, they are very well captured in Figs. 3 and 5, somewhat well in Figs. 4 and 6.

We observe that A_h and A_S behave as if they are spectral complements in the following sense. If smallest eigenvalues of A are well captured by A_h, then the Schur complement is free from smallest eigenvalues (as in Fig. 3), and by (35), so is the preconditioned system (see for example Fig. 5). If not (see Figs. 4, 6), we employ a deflation method as stage two preconditioner on top of M_{left}^{-1}. In Figs. 4 and 6, we illustrate that several smallest eigenvalues of A of varying magnitude are not captured by A_h, they show up as smallest eigenvalues of A_S in varying magnitude. This is

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

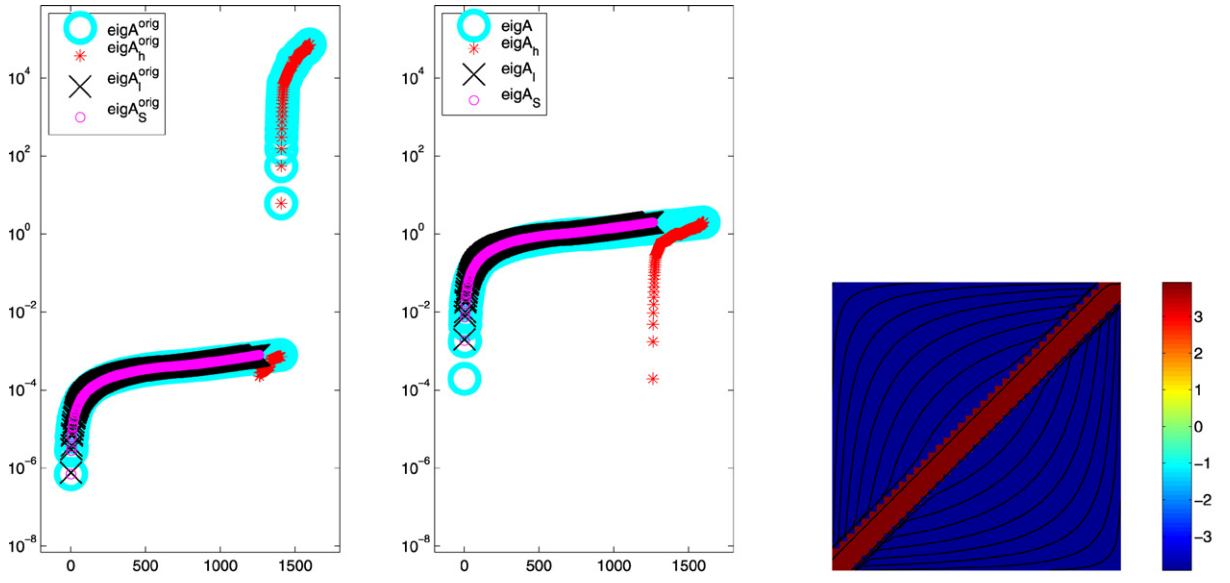

Fig. 3. Test problem 1: Spectra of the subblocks in the original (left) and diagonally scaled (middle) stiffness matrix. (Right) Log permeability field corresponding to a reservoir with one high permeable connected component (island), hence, we observe only one outlier small eigenvalue in the spectrum of the diagonally scaled matrix.

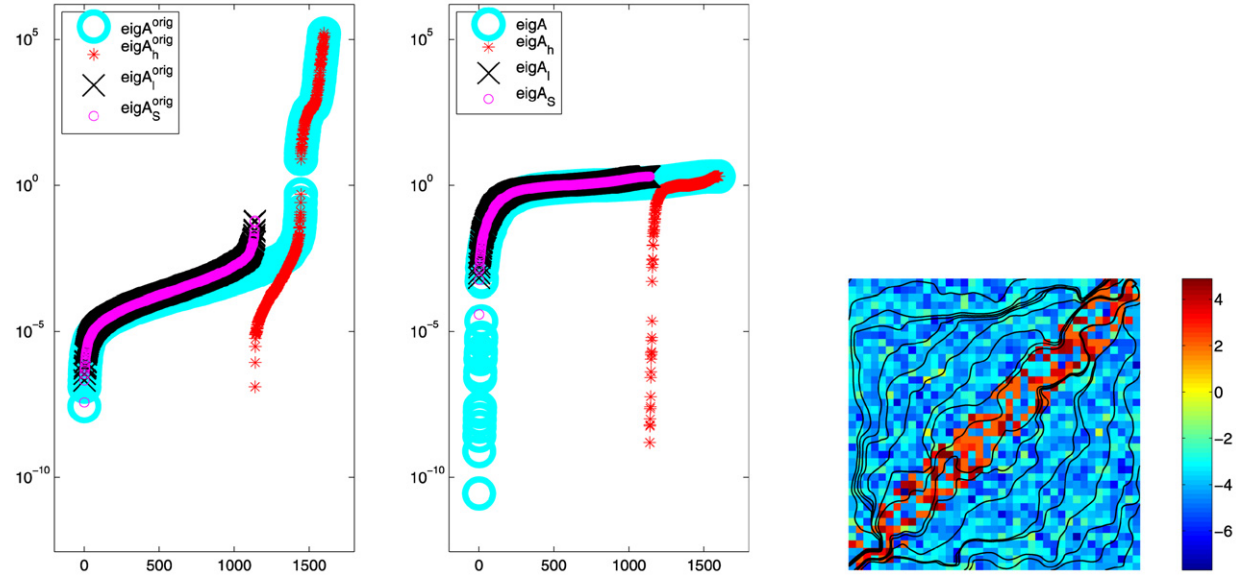

Fig. 4. Test problem 2: The permeability field has a large variation and there are several connected high permeable components, hence, we observe several small eigenvalues of varying magnitude in the spectrum of the diagonally scaled matrix (middle).

exactly where we employ a stage two preconditioner and we show that this strategy is effective as in test problem 6 and Figs. 12 and 13.

To establish a fair comparison among the methods, we stress the implications that the Krylov subspace dimension has for each preconditioner. When a deflation method is employed, the Krylov subspace employed has two pieces. The main piece has $m - l$ vectors and the restarted GMRES uses these vectors. The other piece has l vectors and we extract a near invariant subspace of size l corresponding to the smallest eigenvalues. For instance, $m + l = 50$ means, `harmo, aug, BE` preconditioners are operating on a Krylov subspace of size 40 and the near invariant subspace is extracted by using a collection of $l = 10$ vectors in the harmonic Ritz projec-

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

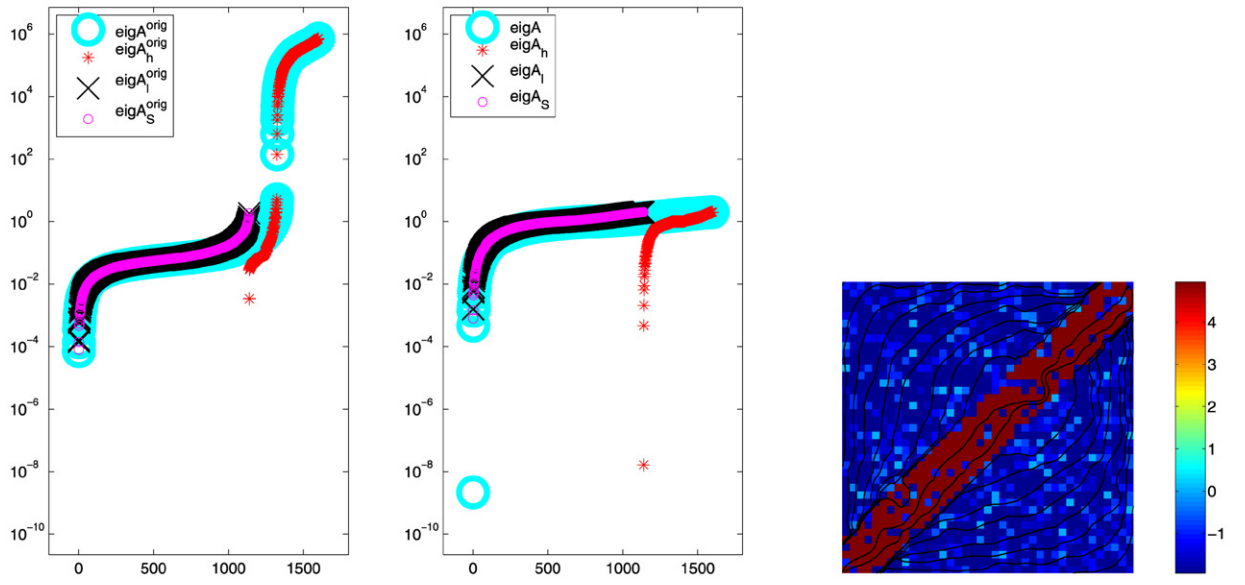

Fig. 5. Test problem 3: The permeability field has a mild variation and the high permeable region forms one big connected component (channel), hence, we observe only one outlier small eigenvalue in the spectrum of the diagonally scaled matrix.

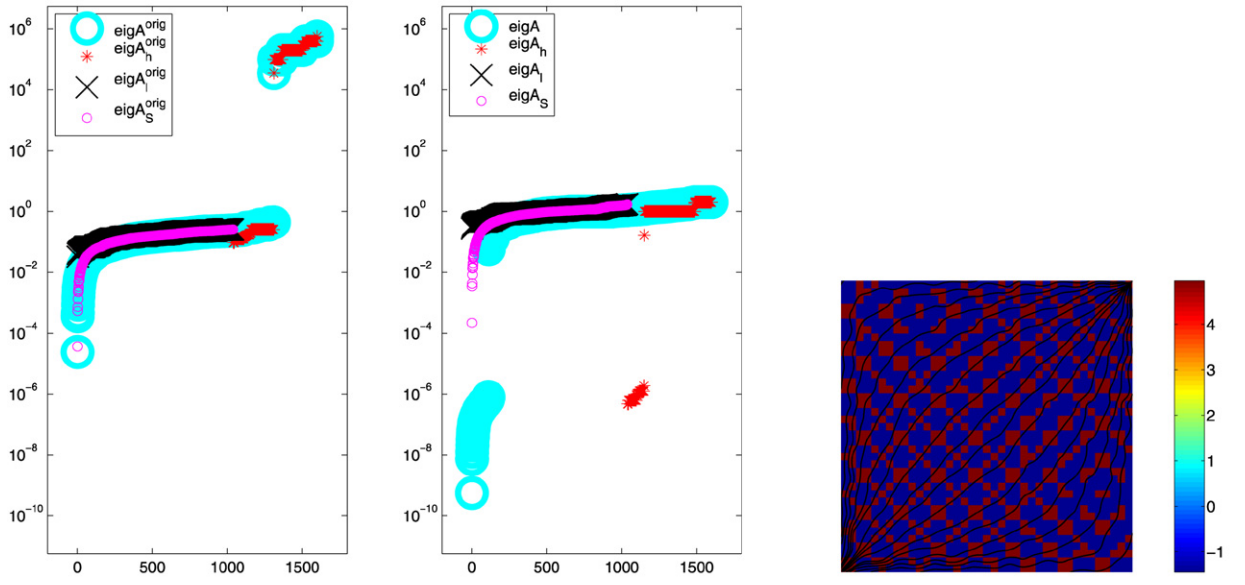

Fig. 6. Test problem 4: The permeability field is checker board like; there are several connected components. When an island touches an other island even at one point, they together form one bigger island. There are several such big islands that create the small eigenvalues in the middle figure.

tion. For all the experiments, the ratio of the Krylov subspace size over the near invariant subspace size is maintained as $m : l = 4 : 1$. Since, deflation preconditioners `2Stage+d(40,10)`, `harmo(40,10)`, `aug(40,10)` and `BE(40,10)` use 50 vectors, we compare them against `gmres(50)`, `2Stage(50)`. In the convergence plots, we always report the case $m + l = 40 + 10$. The iteration count plots contain the following m, l combinations: 8, 2; 12, 3; 16, 4; 20, 5; 24, 6; 28, 7; 32, 8; 36, 9; 40, 10; 44, 11; 48, 12.

The deflation methods we compare are given in Section 4.2. These methods are `harmo(m,l)`, `aug(m,l)`, and `BE(m,l)`. The comparisons are made against `gmres(m+l)`, `2Stage(m+l)`, and `2Stage+d(m,l)` where

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

`2Stage(m+l)` means the two-stage preconditioner without a stage two preconditioner and `2Stage+d(m,l)` means that Algorithm 4.1 is used as a stage two preconditioner.

An important issue is to be able to solve a system in the subblock A_h. We expect to collect smallest eigenvalues in A_h, thereby, this subblock contains the difficult part of the underlying problem. Hence, A_h is ill-conditioned but relatively small in size. In a recursive manner, further ordering can be applied to DOF in A_h if there is extra variation in the high-conductivity values as indicated above. This makes the size of A_h even smaller, hence, its system solve easier. One can utilize various solvers considering the small size of A_h such as direct methods or iterative methods like the deflation methods and AMG. For convenience, the system solve of A_h is done by using the backslash solver in MATLAB.

5.2. Preconditioning results on different test problems

We now proceed to describe in detail the results obtained for each test problem.

5.2.1. Test problem 1

The conductivity field is designed to form a high conductive channel crossing the domain diagonally. Since this channel is sandwiched by low conductive regions, we expect only 1 small eigenvalue in A separated from the rest of the spectrum; see Fig. 3. Indeed, we see an eigenvalue of A of magnitude $\mathrm{O(e-04)}$ which is fully captured by A_h, thereby, $\lambda_{\min}(A_S) = \mathrm{O(e-03)}$ and $\kappa(A_S) = \mathrm{O(e+03)}$.

Deflation methods almost always converged with the exception of `aug(m,l)` and `BE(m,l)` is the most effective. Our preconditioners outperform deflation methods and employing the stage two preconditioner certainly accelerates the convergence; see Figs. 7 and 8. `2Stage+d(40,10)` enjoys the fastest convergence among the methods with $(m, l) = (40, 10)$. The left and right preconditioned versions exhibit similar behavior.

5.2.2. Test problem 2

The conductivity field is designed to test if our preconditioners would work when there is a extreme variability in the coefficients. So both high- and low-conductivity regions contain highly heterogeneous values. Given that $\kappa(A) = \mathrm{O(e+13)}$ and that there is a cluster of smallest eigenvalues of A which ranges between $\mathrm{O(e-10)}$ to $\mathrm{O(e-05)}$, the system matrix is the most difficult among the test problems. The main feature that causes these difficulties is the highly heterogeneous field. Employing M_{left}^{-1} brings $\kappa(A)$ down to $\kappa(M_{\text{left}}^{-1}A) = \mathrm{O(e+07)}$. However, A_h cannot capture the smallest eigenvalues of A, thereby, an eigenvalue of $\mathrm{O(e-05)}$ is formed in A_S and that seems to be the main reason for the convergence failure for `2Stage(50)`. Employing the stage two preconditioner does not seem to eliminate the complication. We observe a reduction in the norm of the error of `2Stage+d(40,10)` whereas all the other methods fail. Left preconditioned `2Stage+d(40,10)` seems to reduce the error more than the right version. In summary, none of the methods can solve this extremely hard problem. This motivates the need to extend the present coefficient separation criteria to capture more complicated conductivity patterns that are not obvious from direct inspection of the permeability field. Multiscale solution methods based on domain decomposition methods [22] or algebraic multigrid [54] could be effective to complement the present approach (either for replacing the role of deflation in M_d or for relaxing the construction of the block A_h). A more reliable assessment of conductive paths can be also obtained from streamline or percolative solutions but with the price of an additional computational penalty.

5.2.3. Test problem 3

This test problem is designed in such a way that the heterogeneity in the conductivity field of test problem 4 is decreased. This decreases the condition number, $\kappa(A) = \mathrm{O(e+10)}$, and only one eigenvalue of $\mathrm{O(e-09)}$ appears in $\sigma(A)$. One can still say that test problem 3 is among the difficult ones.

A_h is very successful in capturing the smallest eigenvalues of A, thereby, $\lambda_{\min}(A_S) = \mathrm{O}(e-03)$ and $\kappa(A_S) = \mathrm{O(e+04)}$. For the left preconditioned case, `2Stage(50)` converges and `2Stage+d(40,10)` converges faster. `2Stage+d(m,l)` converges for a bigger set of (m, l) which makes it more robust with respect to the sizes of the Krylov subspace and the near invariant subspace. Among the deflation methods, `BE(m,l)` seems to be the only converging one. For the left preconditioned case, the error reduction for `2Stage+d(40,10)` is better than that of `BE(40,10)`. We notice that left preconditioning is more effective and results in more convergence when stage two preconditioner is employed.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

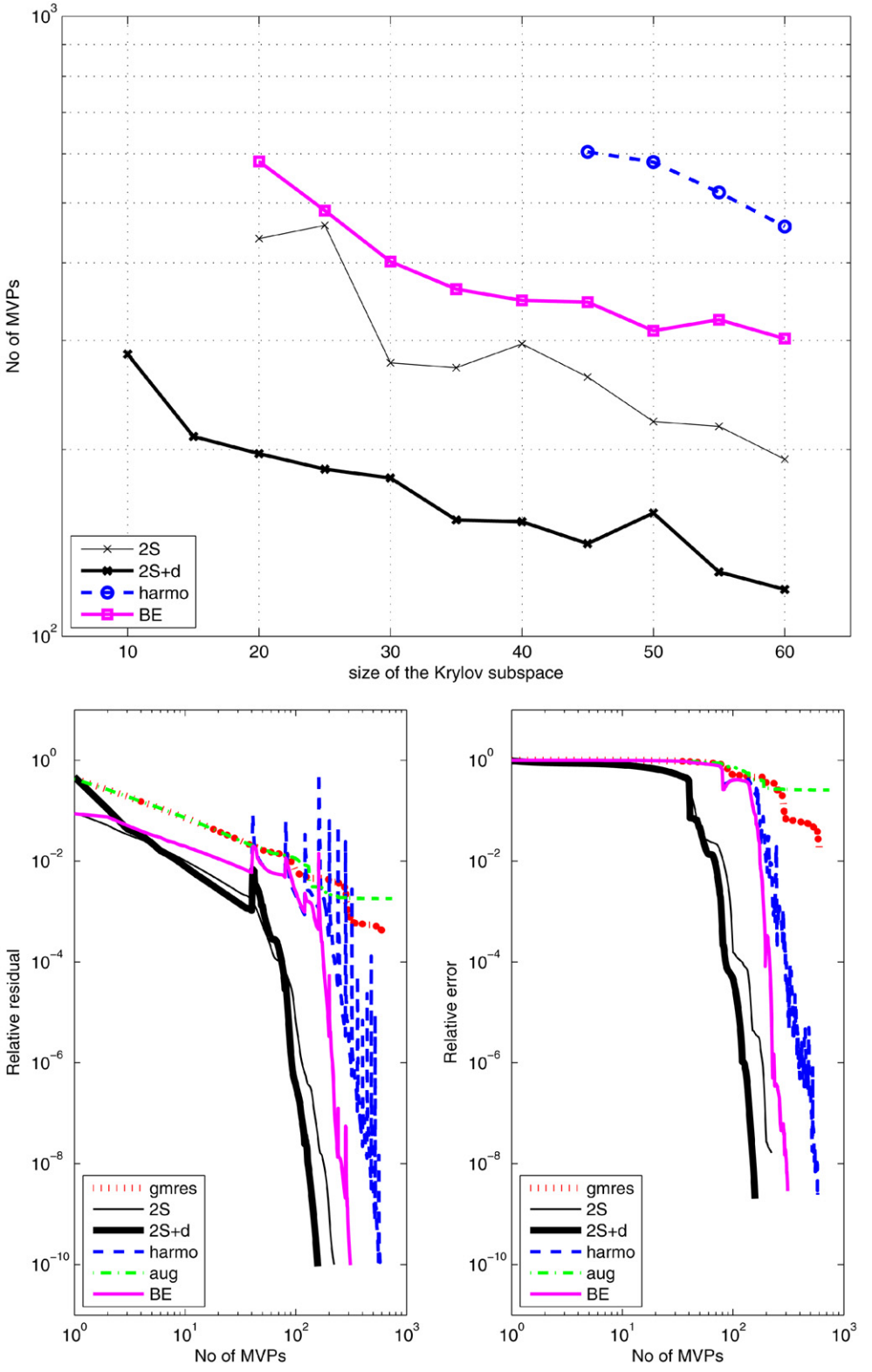

Fig. 7. Test problem 1: (top) comparison of the number of MVPs for different converging m, l combinations. (Bottom) convergence history of the left preconditioner for $m, l = 40, 10$ and $m + l = 40 + 10$.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

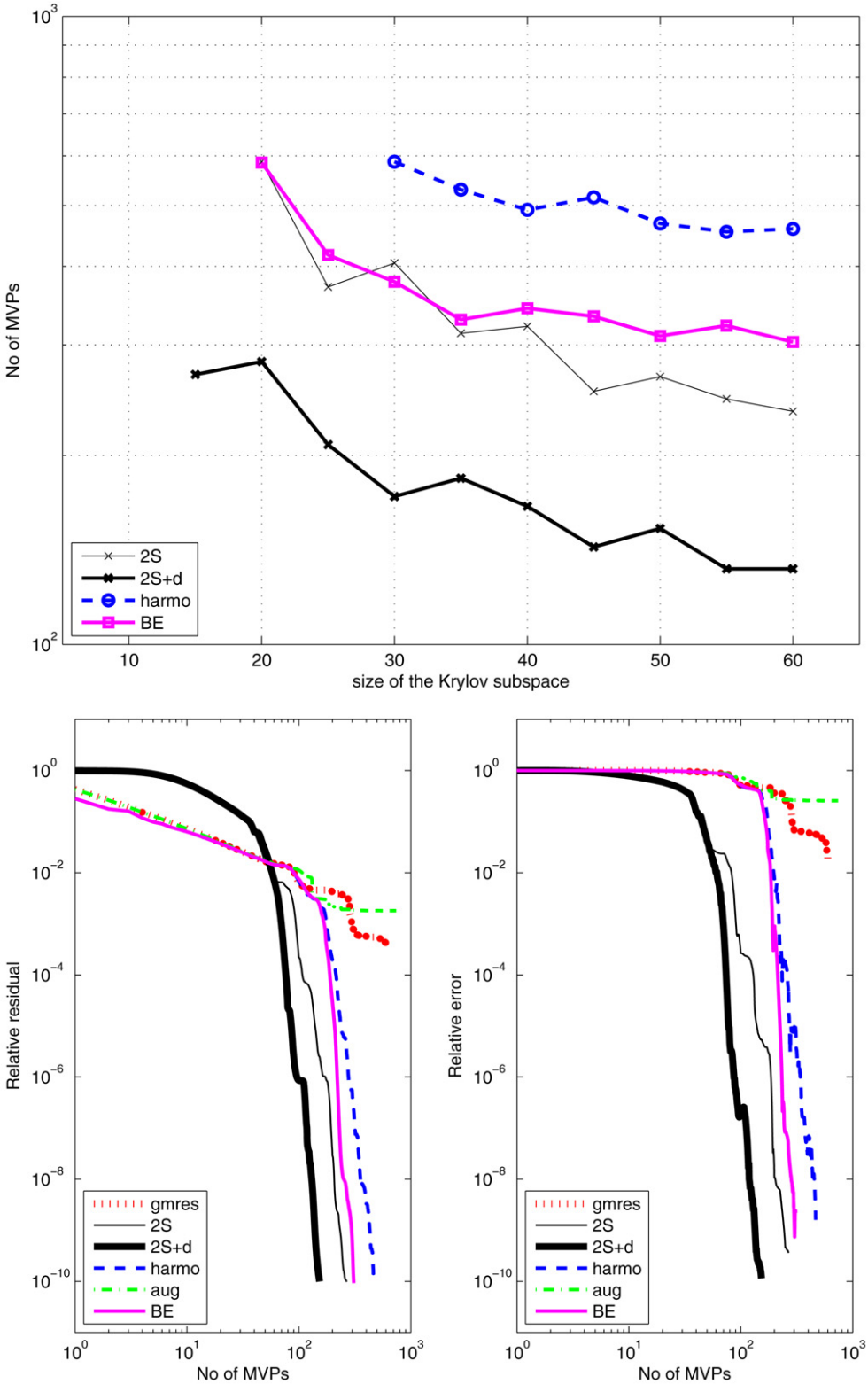

Fig. 8. Test problem 1: (top) comparison of the number of MVPs for different converging m, l combinations. (Bottom) convergence history of the right preconditioner for $m, l = 40, 10$ and $m + l = 40 + 10$.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

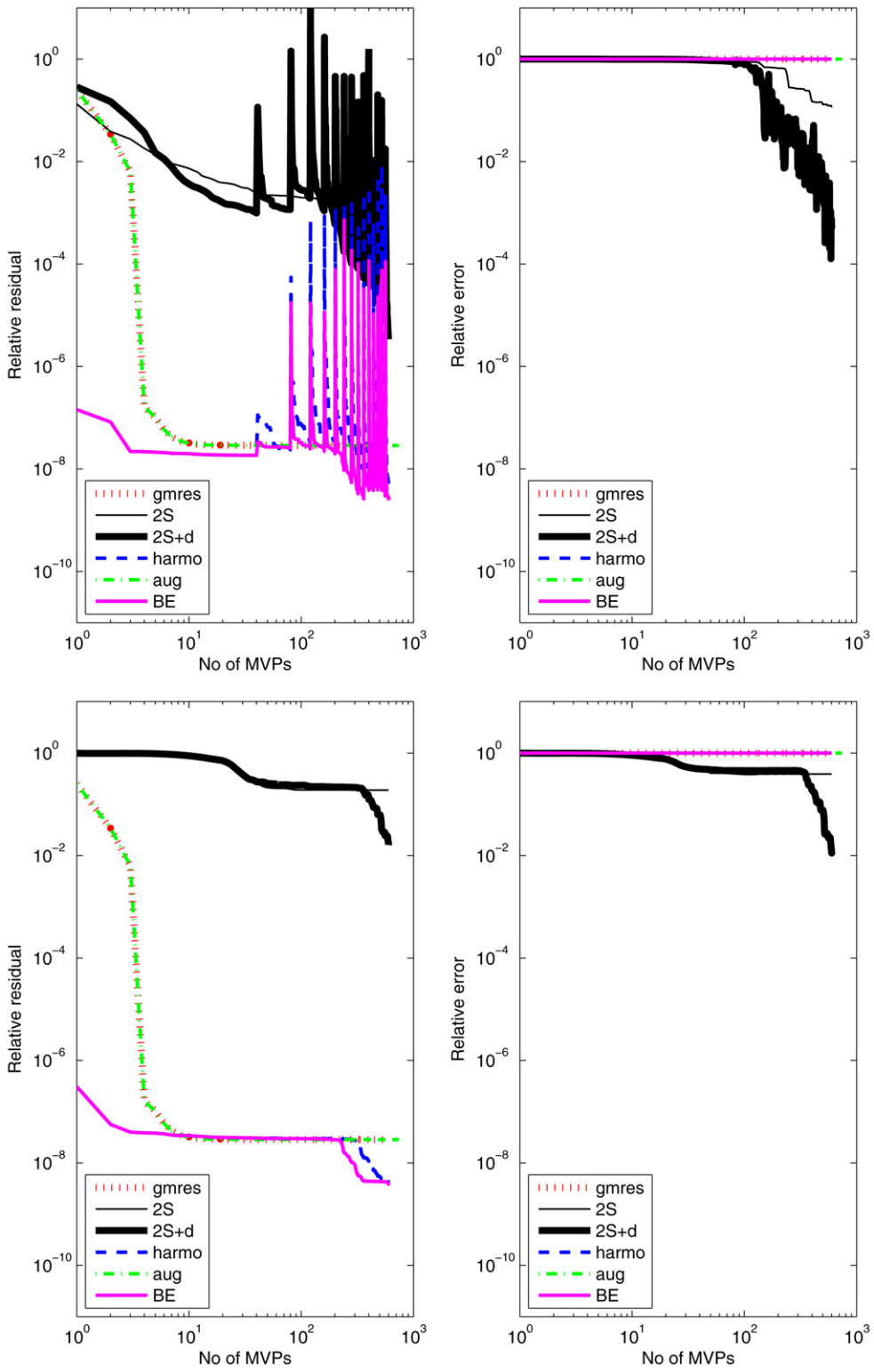

Fig. 9. Test problem 2: convergence history of the left preconditioner (top) and the right preconditioner (bottom) for $m, l = 40, 10$ and $m + l = 40 + 10$.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

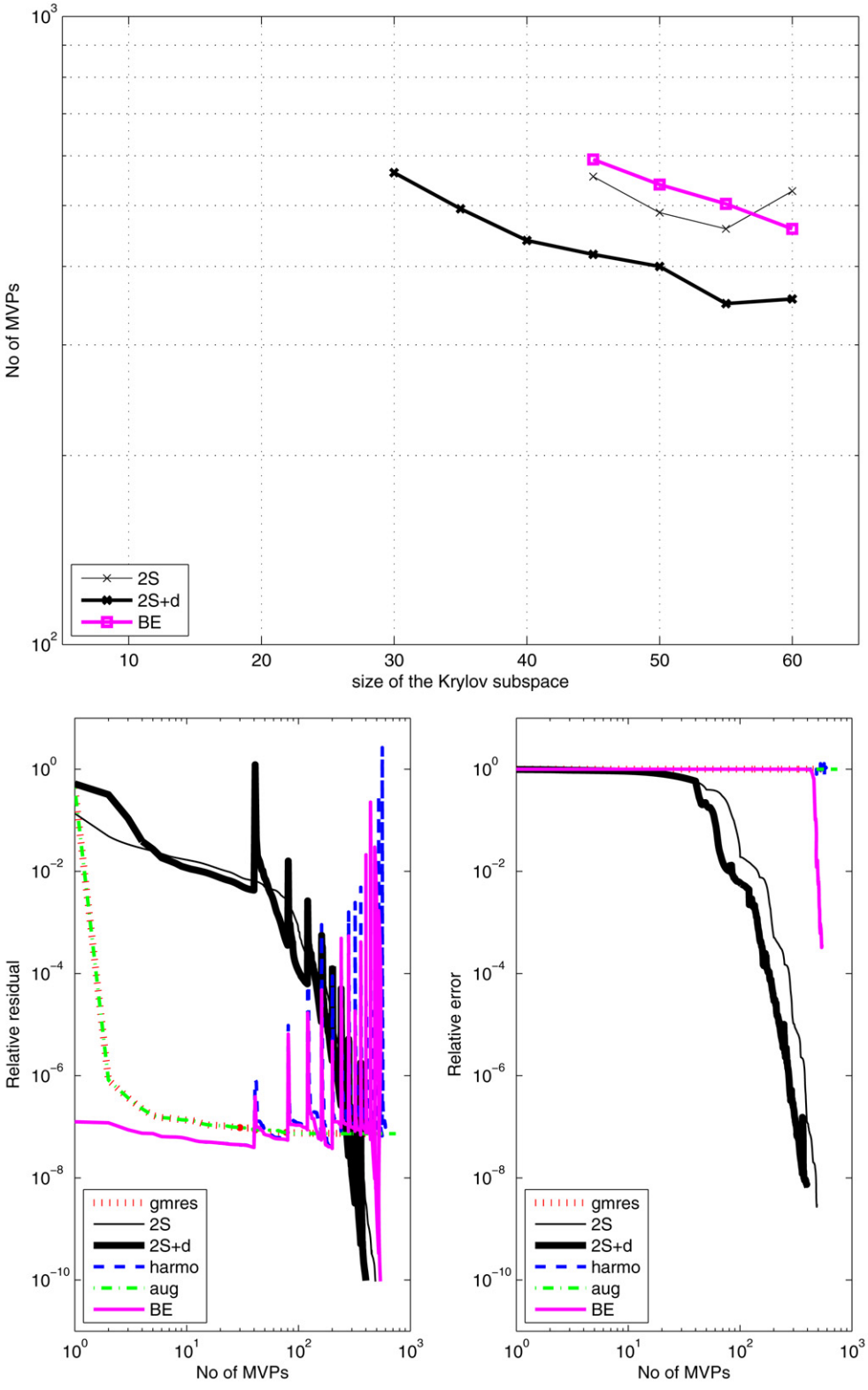

Fig. 10. Test problem 3: (top) comparison of the number of MVPs for different converging m, l combinations. (Bottom) convergence history of the left preconditioner for $m, l = 40, 10$ and $m + l = 40 + 10$; note that only `2Stage(50), 2Stage+d(40,10), BE(40,10)` converge.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

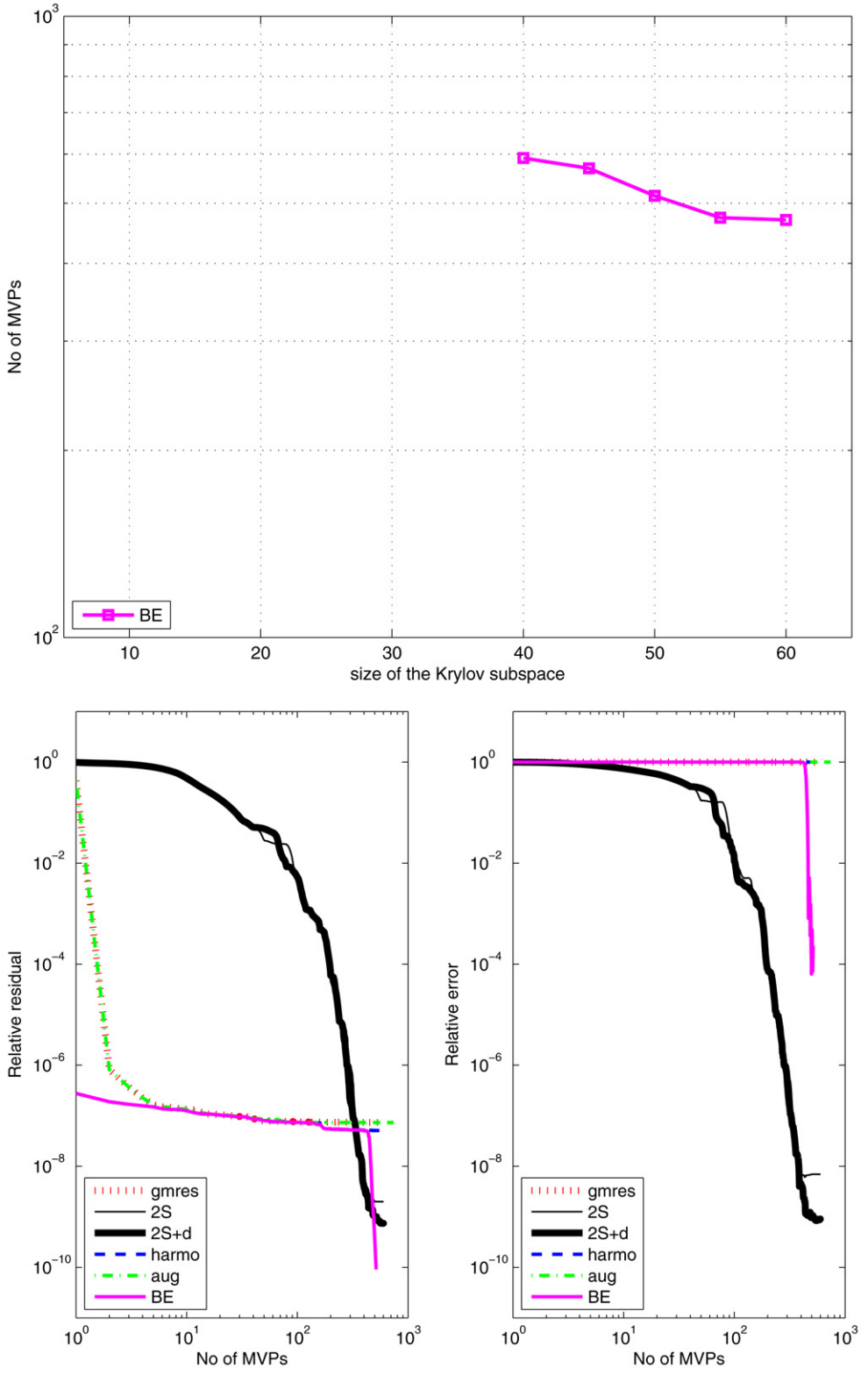

Fig. 11. Test problem 3: (bottom) Convergence history of the right preconditioner for $m, l = 40, 10$ and $m + l = 40 + 10$; note that error reduction for `2Stage(50), 2Stage+d(40,10)` is much better than `BE(40,10)`. However, `2Stage` and `2Stage+d` methods stall and `BE` reaches the stopping tolerance (top).

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

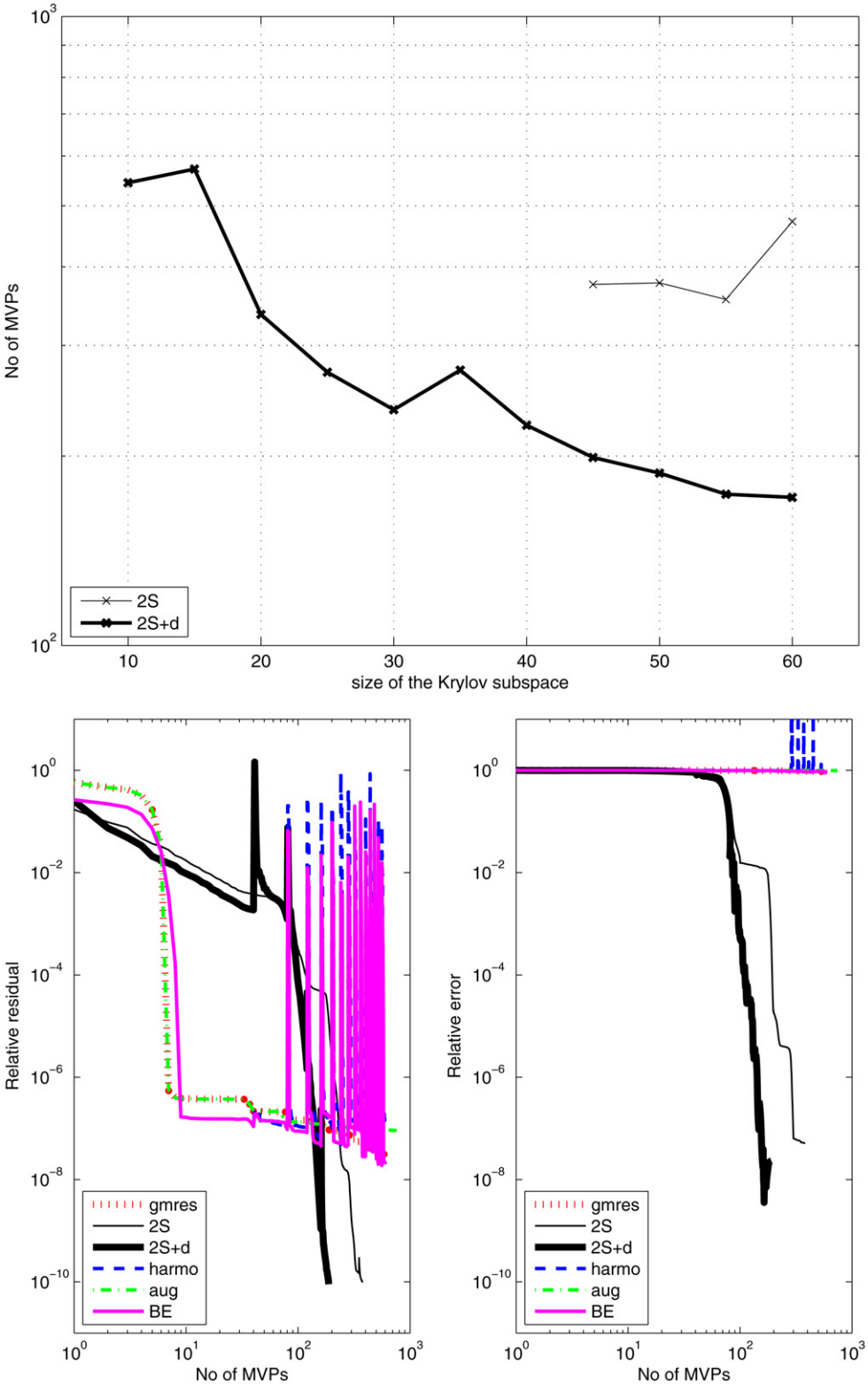

Fig. 12. Test problem 4: (top) comparison of the number of MVPs for different converging m, l combinations. (Bottom) convergence history of the left preconditioner for $m, l = 40, 10$ and $m + l = 40 + 10$. Note that only `2Stage(50), 2Stage+d(40,10)` converge.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

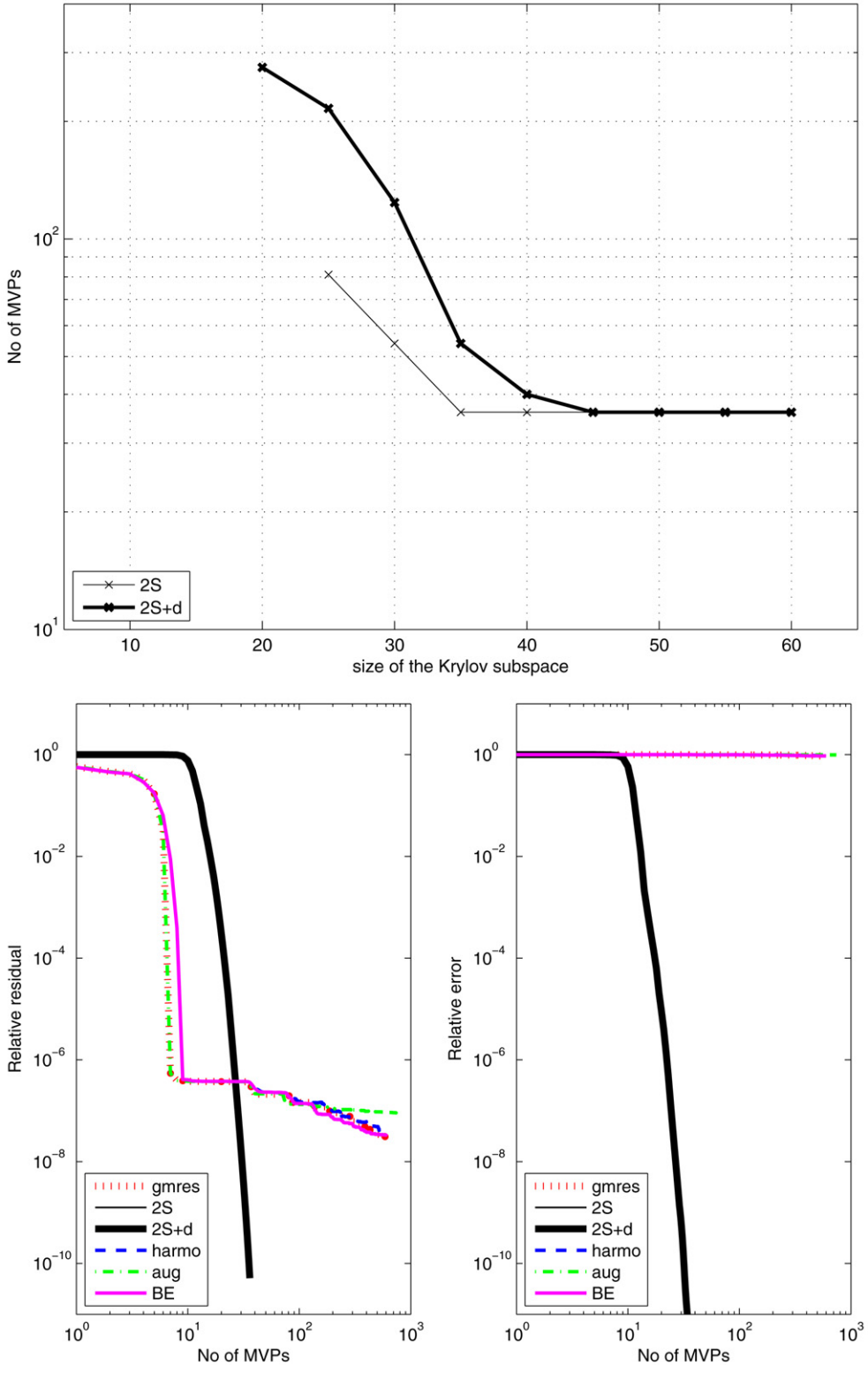

Fig. 13. Test problem 4: (top) comparison of the number of MVPs for different converging m, l combinations. (Bottom) convergence history of the right preconditioner for $m, l = 40, 10$ and $m + l = 40 + 10$. Note that only `2Stage(50), 2Stage+d(40,10)` converge.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

5.2.4. Test problem 4

This test problem contains a conductivity field whose topology resembles a checker board. One can still identify layer-like channels between the opposite corners. This test problem is designed to test if our heuristics about the smallest eigenvalues is valid in this extreme case. Namely, when there is layer-like channels, we can identify contrasting layers, hence, smallest eigenvalues arise due to these contrasts.

Test problem 4 is a difficult one with $\kappa(A) = \mathrm{O(e+10)}$ and A has a cluster of smallest eigenvalues which ranges between $\mathrm{O(e-09)}$ to $\mathrm{O(e-06)}$. Among the smallest eigenvalues, there is only one outlier with magnitude $\mathrm{O(e-09)}$ and A_h cannot capture that. This is reflected to A_S as an outlier eigenvalue where $\lambda_{\min}(A_S) = \mathrm{O(e-04)}$. However, A_h captures smallest eigenvalues of A of $\mathrm{O(e-06)}$ which gives rise to a favorable condition number; $\kappa(M_{\text{left}}^{-1}A) = \mathrm{O(e+04)}$.

We observe a very effective `2Stage+d(m,l)` for both left and right preconditioned cases whereas all the deflation methods fail. Furthermore, in the left preconditioned case, `2Stage(50)` converges but slower than `2Stage+d(40,10)`. `2Stage(m+l)` does not converge for most of (m, l) combinations, whereas employing a stage two preconditioner brings robustness and `2Stage+d(m,l)` converges for almost all (m, l) combinations in both left and right preconditioned cases. This is an effective use of the stage two feature of our preconditioner which addresses the smallest eigenvalues of A_S. In the left preconditioned case, we notice a more improved convergence when the stage two preconditioner is employed compared to right preconditioned case. However, right preconditioner gives a constant number of 36 iterations for a large combination of (m, l) values where as left preconditioner iteration counts are in the vicinity of 400.

In all cases, we observe in the numerical experiments that the performance of left and right preconditioners are similar. The left preconditioner as explained in Section 3.2, can create increasing residual at the GMRES(m) restarts due to the updated near invariant subspace.

6. Conclusions

In this article, we present two-stage physics-based preconditioners that are designed to address severe contrasts in diffusion coefficients. The contrasts give rise to extremely small eigenvalues and they seem to be the main bottleneck for iterative solvers. The main assumption behind this work is that a highly connected conductivity network strongly defines and governs diffusive response throughout a material. Hence, conductivity should dictate the way we solve linear equations. To that end, we propose a two-stage physics based preconditioner with an optional stage two deflation preconditioner as outlined in Algorithm 3.1.

Since, historically deflation methods are designed to address extremal eigenvalues, we compare our preconditioner to these methods as well as use them as an optional stage two preconditioner. We compare our preconditioner to three well-known deflation methods: harmonic [17], augmented [34], and Burrage and Erhel [8]. These are *dynamic* deflation methods where the near invariant subspace U is extracted by a harmonic Ritz projection from the Hessenberg matrix. This is the standard way of computing the near invariant subspaces in all of our numerical experiments. We report that our preconditioners – even without the stage two preconditioner – outperform all of the three methods. The best competitor and the most unstable methods are `Burrage-Erhel(m,l)` and `augment(m,l)`, respectively. Furthermore, our preconditioners are robust with respect to the Krylov subspace size. We report convergence for far more combinations of (m, l) compared to deflation methods.

The selection of DOF in A_h is purely algebraic, hence, we emphasize that our preconditioners are quite flexible to be adapted for any discretization method. The main advantage of our preconditioners is the fact that they can handle *flexible* and *realistic* geometries. The performance and computational cost of the left and right preconditioners are similar. However, the left preconditioner can create increasing residual at the GMRES(m) restarts due to the updated near invariant subspace.

The condition number of the linear system depends both on the mesh size Δx and the coefficient size m. We assume a fixed Δx and address only the m dependence. The Δx dependence will be investigated in the companion article [3]. As $m \to \infty$, we showed that the condition number of the preconditioned system becomes independent of m. Namely, the preconditioner eliminates the m dependence. After our preconditioner is applied, in the limiting case, the remaining step is to construct an effective preconditioner for the Schur complement with respect to Δx only which indicates that there is a decoupling of m and Δx.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

Despite the effectiveness of the proposed physics-based preconditioners there are many research issues that remain open. We list some of the one that we consider promising to address in the near future:

- In some stringent situations, deflation may not be sufficiently robust as a second stage preconditioner in the advent of either extreme ill-posedness (due to inner significant heterogeneities in low- or high-conductive regions) or to the size of the block A_l. Thus, we need to explore variations of the method such as multi-stage preconditioning and the use of solvers such as algebraic multigrid (AMG), algebraic multilevel iterations (AMLI) and sparse approximate inverse (SPAI) methods to hopefully speedup the solution of both the A_h and A_l blocks.
- Our method is based on static conductivity information and does not account for a better assessment of the true media connectivity. In view of the results obtained from the stringent test problem 2, we believe that streamlines or percolative methods should be very useful in defining improved physics-based preconditioning strategies since they are designed to detect preferential flow paths in a more reliable way.
- Connections of nested versions of the present methodology (i.e., multi-stage preconditioning) with AMG methods seems to be in order to design improve solution heuristics according to the connectivity strength between matrix coefficients and the underlying physical domain. These issues will be addressed in the upcoming article [4].
- Last but not least, many researchers are devoted to the development of accurate and efficient multiscale methods. We believe that our method can be seen as high-level approach where these methods can be further extended to tackle fine scale solutions such as in [1,22,26].

Acknowledgements

The authors want to thank Ivan Graham, Robert Scheichl, and Jan Van Lent from the University of Bath at UK, for the interesting discussions on the addressed topic during their visit to CSM.

References

[1] J.E. Aarnes, On the use of a mixed multiscale finite element method for greater flexibility and increased speed or improved accuracy in reservoir simulation, Multiscale Model. Simul. 2 (3) (2004) 421–439.
[2] J.E. Aarnes, T.Y. Hou, Multiscale domain decomposition methods for elliptic problems with high aspect ratios, Acta Math. Appl. Sinica 18 (1) (2002) 63–76.
[3] B. Aksoylu, I.G. Graham, H. Klie, R. Scheichl, Towards a rigorously justified algebraic preconditioner for high-contrast diffusion problems, Comput. Vis. Sci. 4–6 (11) (2008) 319–331.
[4] B. Aksoylu, H. Klie, Two-stage percolation aggregation preconditioners for exploiting reservoir connectivity and heterogeneity, Computational Geociences, 2008, submitted for publication.
[5] B. Aksoylu, H. Klie, M.F. Wheeler, Physics-based preconditioners for porous media flow applications. Technical report, The University of Texas at Austin, Institute for Computational Engineering and Sciences, Austin, TX, ICES Report 07-08, April, 2007, http://www.ices.utexas.edu/research/reports/2007/0708.pdf.
[6] O. Axelsson, Iterative Solution Methods, Cambridge University Press, 1994.
[7] L. Blank, Preconditioning via a Schur complement method: An application in state estimation, SIAM J. Sci. Comput. 25 (3) (2003) 942–960.
[8] K. Burrage, J. Erhel, On the performance of various adaptive preconditioned GMRES strategies, Numer. Linear Algebra Appl. 5 (1998) 101–121.
[9] C. Le Calvez, B. Molina, Implicit restarted and deflated GMRES, Numer. Algebra 21 (1999) 261–285.
[10] H. Cao, H.A. Tchelepi, J. Wallis, H. Yardumian, Parallel scalable unstructured CPR-type linear solver for reservoir simulation, in: Society of Petroleum Engineers, SPE Annual Technical Conference and Exhibition, Dallas, TX, 9–12 October 2005.
[11] T.F. Chan, Deflated decomposition of solutions of nearly singular systems, SIAM J. Sci. Stat. Comput. 5 (1) (1984) 121–134.
[12] T.F. Chan, Deflation techniques and block-elimination algorithms for solving bordered singular systems, SIAM J. Numer. Anal. 21 (4) (1984) 738–754.
[13] A. Chapman, Y. Saad, Deflated and augmented Krylov subspace techniques, Numer. Linear Algebra Appl. 4 (1997) 43–66.
[14] K. Chen, Matrix Preconditioning Techniques and Applications, Cambridge University Press, Cambridge, UK, 2005.
[15] C. Dawson, H.M. Klíe, M.F. Wheeler, C. Woodward, A parallel implicit, cell-centered method for two-phase flow with a preconditioned Newton–Krylov solver, Comp. Geosci. 1 (1997) 215–249.
[16] L.J. Durlofsky, Numerical calculation of equivalent grid block permeability tensors for heterogeneous porous media, Water Resources Res. 27 (5) (1991) 699–708.
[17] J. Erhel, K. Burrage, B. Pohl, Restarted GMRES preconditioned by deflation, J. Comput. Appl. Math. 69 (1996) 303–318.
[18] J. Frank, C. Vuik, On the construction of deflation-based preconditioners, SIAM J. Sci. Comput. 23 (2) (2001) 442–462.
[19] L. Fung, A.H. Dogru, Parallel unstructured solver methods for complex giant reservoir simulation, in: Society of Petroleum Engineers, SPE Reservoir Simulation Symposium, Houston, TX, Feb. 26–28, 2007.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002

[20] S. Goossen, D. Roose, Ritz and harmonic Ritz values and the convergence of FOM and GMRES, Numer. Linear Algebra Appl. 6 (1999) 281–293.
[21] I.G. Graham, M.J. Hagger, Unstructured additive Schwarz-conjugate gradient method for elliptic problems with highly discontinuous coefficients, SIAM J. Sci. Comput. 20 (6) (1999) 2041–2066.
[22] I.G. Graham, P. Lechner, R. Scheichl, Domain decomposition for multiscale PDEs, Technical Report Preprint 11/06, Bath Institute For Complex Systems, University of Bath, UK, 2006, available at http://www.bath.ac.uk/math-sci/BICS.
[23] L. Greengard, J.-Y. Lee, Electrostatics and heat conduction in high contrast composite materials, J. Comput. Phys. 211 (2006) 64–76.
[24] H. Waisman, J. Fish, R.S. Tuminaro, J. Shadid, The generalized global basis (GBB) method, Int. J. Numer. Meth. Engng. 61 (2004) 1243–1269.
[25] N.J. Higham, Accuracy and Stability of Numerical Algorithms, SIAM, Philadelphia, USA, 2002.
[26] T.Y. Hou, X.H. Wu, A multiscale finite element method for elliptic problems in composite materials and porous media, J. Comput. Phys. 134 (1997) 169–189.
[27] H. Klie, Krylov–Secant methods for solving large scale systems of coupled nonlinear parabolic equations, PhD thesis, Dept. of Computational and Applied Mathematics, Rice University, Houston, TX, 1996.
[28] K. Lust, D. Roose, A. Spence, A.R. Champneys, An adaptive Newton–Picard algorithm with subspace iteration for computing periodic solutions, SIAM J. Sci. Comput. 19 (1998) 1188–1209.
[29] M.S. Lynn, W.P. Timlake, The use of multiple deflations in the numerical solution of singular systems of equations, with applications to potential theory, SIAM J. Numer. Anal. 5 (2) (1968) 303–322.
[30] S.P. MacLachlan, J.D. Moulton, Multilevel upscaling through variational coarsening, Water Resources Res. 42 (2006) 1–9.
[31] L. Mansfield, On the conjugate gradient solution of the Schur complement system obtained from domain decomposition, SIAM J. Numer. Anal. 27 (1990) 1612–1620.
[32] L. Mansfield, Damped Jacobi preconditioning and coarse grid deflation for conjugate gradient iteration on parallel computers, SIAM J. Sci. Stat. Comput. 12 (6) (1991) 1314–1323.
[33] J. Monteagudo, A. Firoozabadi, Numerical simulation of water injection in disconnected and connected fractured media using Jacobian-free fully implicit control volume method, in: 14th Symposium on Improved Oil Recovery. SPE paper No. 89449, Tulsa, OK, 2004.
[34] R.B. Morgan, A restarted GMRES method augmented with eigenvectors, SIAM J. Matrix Anal. Appl. 16 (4) (1995) 1154–1171.
[35] R.B. Morgan, Implicit restarted GMRES and Arnoldi methods for nonsymmetric systems of equations, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1112–1135.
[36] R.B. Morgan, GMRES with deflated restarting, SIAM J. Sci. Statist. Comput. 24 (1) (2002) 20–27.
[37] R.B. Morgan, Restarted block-GMRES with deflation of eigenvalues, Appl. Numer. Math. 28 (2004) 1–15.
[38] R. Nabben, C. Vuik, A comparison of deflation and coarse grid correction applied to porous media flow, SIAM J. Numer. Anal. 42 (2004) 1631–1647.
[39] R.A. Nicolaides, Deflation of conjugate gradients with applications to boundary value problems, SIAM J. Numer. Anal. 24 (1987) 355–365.
[40] V.A. Nœtinger, V. Artus, G. Zargar, The future of stochastic and upscaling methods in hydrogeology, Hydrogeology J. 13 (2005) 184–201.
[41] T.L. Noorden, S.M. Verduyn, A. Bliek, A Broyden rank $p + 1$ update continuation method with subspace iteration, SIAM J. Sci. Comput. 25 (2004) 1921–1940.
[42] P. Renard, G. Le Loch, E. Ledoux, G. de Marsily, R. Mackay, A fast algorithm for the estimation of the equivalent hydraulic conductivity of heterogeneous media, Water Resources Res. 36 (12) (2000) 3567–3580.
[43] P. Renard, G. Le Loch, E. Ledoux, G. de Marsily, R. Mackay, On the use of apparent hydraulic diffusivity as an indicator of connectivity, J. Hydrology 329 (2006) 377–389.
[44] W.R. Rheinboldt, Numerical methods for a class of finite dimensional bifurcation problems, SIAM J. Numer. Anal. 15 (1977) 1–11.
[45] M.J. Ronayne, S.M. Gorelick, Effective permeability of porous media containing branching channel networks, Phys. Rev. E 73 (026305) (2006) 1–10.
[46] T.F. Russell, M.F. Wheeler, Finite element and finite difference methods for continuous flows in porous media, in: R.E. Ewing (Ed.), The Mathematics of Reservoir Simulation, SIAM, Philadelphia, USA, 1983, pp. 35–106.
[47] Y. Saad, Analysis of augmented Krylov subspace methods, SIAM J. Matrix Anal. Appl. 18 (2) (1997) 435–449.
[48] Y. Saad, Iterative Methods for Sparse Linear Systems, second ed., SIAM, Philadelphia, USA, 2003.
[49] Y. Saad, M. Yeung, J. Erhel, F. Guyomar'h, A deflated version of the conjugate gradient algorithm, SIAM J. Sci. Comput. 21 (5) (2000) 1909–1926.
[50] X. Sanchez-Vila, A. Guadagnini, J. Carrera, Representative hydraulic conductivities in saturated groundwater flow, Rev. Geophys. 44 (2006) 1–46.
[51] R. Scheichl, R. Masson, J. Wendebourg, Decoupling and block preconditioning for sedimentary basin formulations, Comput. Geosci. 7 (2003) 295–318.
[52] V. Shenoy, Multi-scale modeling strategies in materials science and the quasicontinuum method, Bull. Mater. Sci. 26 (1) (2003) 53–62.
[53] G.W. Stewart, On the implicit deflation of nearly singular systems of linear equations, SIAM J. Sci. Stat. Comput. 2 (2) (1981) 136–140.
[54] K. Stuben, T. Kees, H. Klie, M.F. Wheeler, Algebraic Multigrid Methods (AMG) for the efficient solution of fully implicit formulations in reservoir simulation, in: SPE Reservoir Simulation Symposium, Houston, TX, Feb. 26–28, 2007. SPE paper No. 105832.
[55] F. Vermolen, C. Vuik, A. Segal, Deflation in preconditioned conjugate gradient methods for finite element problems, in: Conjugate Gradient and Finite Element Methods, Springer-Verlag, Berlin, 2004, pp. 103–129. SCS.
[56] C. Vuik, A. Segal, J.A. Meijerink, An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts of coefficients, J. Comput. Phys. 152 (1999) 385–403.
[57] C. Vuik, A. Segal, J.A. Meijerink, G.T. Wijma, The construction of projection vectors for a ICCG method applied to problems with extreme contrasts in the coefficients, J. Comput. Phys. 172 (2001) 426–450.
[58] T.I. Zohdi, J.T. Oden, G.J. Rodin, Hierarchical modeling of heterogeneous bodies, Comput. Meth. Appl. Mech. Eng. 138 (1996) 273–298.

Please cite this article in press as: B. Aksoylu, H. Klie, A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media, Applied Numerical Mathematics (2008), doi:10.1016/j.apnum.2008.06.002