Saddle point preconditioners for weak-constraint 4D-Var

Electronic Transactions on Numerical Analysis.
Volume XX, pp. 1–yy, 20xx.

ISSN 1068–9613.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

SADDLE POINT PRECONDITIONERS FOR WEAK-CONSTRAINT 4D-VAR

JEMIMA M. TABEART* AND JOHN W. PEARSON†

Abstract. Data assimilation algorithms combine information from observations and prior model information to obtain the most likely state of a dynamical system. The linearised weak-constraint four-dimensional variational assimilation problem can be reformulated as a saddle point problem, which admits more scope for preconditioners than the primal form. In this paper we design new terms which can be used within existing preconditioners, such as block diagonal and constraint-type preconditioners. Our novel preconditioning approaches: (i) incorporate model information, and (ii) are designed to target correlated observation error covariance matrices. To our knowledge (i) has not previously been considered for data assimilation problems. We develop new theory demonstrating the effectiveness of the new preconditioners within Krylov subspace methods. Linear and non-linear numerical experiments reveal that our new approach leads to faster convergence than existing state-of-the-art preconditioners for a broader range of problems than indicated by the theory alone. We present a range of numerical experiments performed in serial.

Key words. Saddle point systems, Variational data assimilation, Preconditioning

AMS subject classifications. 65F08, 65F10, 65N21

1. Introduction.

Data assimilation has seen substantial interest in fields such as numerical weather prediction [6, 32], ecology [30, 31], and hydrology [7, 44] in recent decades. The variational data assimilation problem can be written mathematically as follows: For a given time window $[t_0, t_N]$, let $\mathbf{x}_i^t \in \mathbb{R}^s$ be the true state of a dynamical system of interest at time t_i, where s is the number of state variables. Data assimilation algorithms combine observations of a dynamical system, $\mathbf{y}_i \in \mathbb{R}^{p_i}$ at times t_i, with prior information from a model, $\mathbf{x}_b \in \mathbb{R}^s$, to find $\mathbf{x}_i \in \mathbb{R}^s$, the most likely state of the system at time t_i. The prior, or background state, is valid at initial time t_0 and can be written as an approximation to the true state via $\mathbf{x}_b = \mathbf{x}_0^t + \epsilon^b$. We assume that the background errors $\epsilon^b \sim \mathcal{N}(0, \mathbf{B})$, where $\mathbf{B} \in \mathbb{R}^{s\times s}$ is a background error covariance matrix. In order to compare observations made at different locations, or of different variables to those in the state vector $\mathbf{x}_i$, we define a, possibly non-linear, observation operator $\mathcal{H}_i : \mathbb{R}^s \to \mathbb{R}^{p_i}$ which maps from state variable space to observation space at time t_i. Observations at time t_i are written as $\mathbf{y}_i = \mathcal{H}_i[\mathbf{x}_i^t] + \epsilon_i \in \mathbb{R}^{p_i}$, for $i = 0, 1, \ldots, N$, where the observation error $\epsilon_i \sim \mathcal{N}(0, \mathbf{R}_i)$ and $\mathbf{R}_i \in \mathbb{R}^{p_i\times p_i}$ are observation error covariance matrices.

In weak-constraint four-dimensional varational data assimilation (4D-Var) the state $\mathbf{x}_{i-1}$ at time t_{i-1} is propagated to the next observation time t_i using an imperfect forecast model, $\mathcal{M}_i$, to obtain $\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1}) + \epsilon_i^m$. The model error at each time is given by $\epsilon_i^m \sim \mathcal{N}(0, \mathbf{Q}_i)$, where $\mathbf{Q}_i \in \mathbb{R}^{s\times s}$ is the model error covariance matrix at time t_i. It is typically assumed that the error covariance matrices are mutually uncorrelated across different types and different observation times. The analysis, or most likely state $\mathbf{x}_0$ at time t_0, minimises the weak constraint 4D-Var objective function given by

$$J(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N) = (\mathbf{x}_0 - \mathbf{x}_b)^\top \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \sum_{i=0}^{N}(\mathbf{y}_i - \mathcal{H}_i[\mathbf{x}_i])^\top \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathcal{H}_i[\mathbf{x}_i])$$
$$+ \sum_{i=1}^{N}(\mathbf{x}_i - \mathcal{M}_i(\mathbf{x}_{i-1}))^\top \mathbf{Q}_i^{-1}(\mathbf{x}_i - \mathcal{M}_i(\mathbf{x}_{i-1})). \tag{1.1}$$

*Department of Mathematics and Computer Science, Eindhoven University of Technology, De Zaale, Eindhoven, 5612 AZ, The Netherlands (j.m.tabeart@tue.nl)

†School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom (j.pearson@ed.ac.uk)

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Weak-constraint 4D-Var is used in numerical weather prediction (NWP) to estimate the initial condition for a weather forecast [42, 43]. Practical implementations of (1.1) pose mathematical and computational challenges. Firstly, the dimension of the problem can be large: in NWP [6] the dimension of the state can be of order 10^9, and the number of observations can be of order 10^6. This application is also time critical, so the time that can be allocated to the data assimilation procedure is very limited in operational situations. Computational efficiency is therefore vital, and designing techniques to ensure fast convergence of the minimisation of the objective function has been an ongoing area of research interest, see for instance [13, 5, 9, 21].

The weak constraint objective function (1.1) is typically solved via an incremental approach, where a small number of non-linear outer loops and a larger number of linearised inner loops are solved [16, 17]. Standard Krylov subspace solvers can then be applied for the inner loops. An alternative approach involves solving the linearised inner loop using a saddle point formulation [15, 13, 18, 8], which admits a richer choice of preconditioning structures compared to the primal form. Prior work has developed specific preconditioners for the saddle point data assimilation problem [15, 13], typically focusing on approximations to the term containing information about the linearised model. Block diagonal preconditioners are appealing, due to the potential to apply the MINRES algorithm and develop guaranteed theoretical insight about the convergence rates based on the eigenvalues of the preconditioned system, although inexact constraint preconditioners with GMRES have been found to yield better performance in data assimilation settings for a variety of problems [15, 14].

Recent research on the primal form of the variational data assimilation problem has revealed that the observation error covariance matrices $\mathbf{R}_i$ play an important role in determining the convergence of iterative methods [38, 40]. In the last decades, researchers have increasingly made use of observing systems that have correlated observation error covariance matrices [45, 37]. Previous saddle point preconditioners typically applied the exact inverses of $\mathbf{R}_i$ [15, 18], which is known to be computationally infeasible for many satellite observing systems [45, 39]. It is therefore expected that new terms within saddle point preconditioners which incorporate correlated information from $\mathbf{R}_i$ and are inexpensive to apply could be beneficial in terms of convergence for many observing systems. In what follows, we therefore consider preconditioners for the observation error covariance matrix that explicitly allow for full correlation matrices.

In this paper, we consider novel terms within existing preconditioning structures for the saddle point framework, with particular focus on the correlated observation error setting. To our knowledge, our preconditioners account explicitly for model information within the model term for the first time. We begin in Section 2 by introducing the saddle point formulation of the weak-constraint 4D-Var data assimilation problem, and presenting existing state-of-the-art preconditioners for the saddle point setting. In Section 3 we prove bounds on the eigenvalues of a block diagonal preconditioned system in terms of constituent matrices of the saddle point problem. In Sections 4 and 5 we then analyse the effect of applying existing and new preconditioners for the model term and observation error covariance term respectively. In Section 6 we present our numerical framework for the experimental results in Section 7. For both the non-linear Lorenz 96 problem and linear heat equation problem we find that our new preconditioners result in a reduced iteration count compared to block diagonal and inexact constraint preconditioners in a variety of settings, in the presence of correlated observation errors. Although the theoretical guarantees only apply to the block diagonal preconditioner, qualitative behaviour is similar for the inexact constraint preconditioner, with large reductions in iterations and matrix–vector products. Finally, in Section 8 we present our conclusions.

2. Background.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

2.1. Data assimilation. In this section we introduce the saddle point formulation of the weak constraint four-dimensional variational (4D-Var) data assimilation problem (1.1). Given a time window $[t_0, t_N]$, split into N subwindows, we wish to find the state $\mathbf{x} \in \mathbb{R}^s$ at time t_0 that is closest in a weighted norm sense both to the observations throughout the time window, and to prior information at the initial time propagated by a model. The incremental primal formulation updates $\mathbf{x}_0^{(l+1)} = \mathbf{x}_0^{(l)} + \boldsymbol{\delta}\mathbf{x}^{(l)}$ by solving an objective function via a series of inner and outer loops to find a sequence of increments to the background state $\mathbf{x}_b = \mathbf{x}_0^{(0)}$. In the inner loop a linearised problem is solved, typically using iterative solvers such as the Conjugate Gradient method [23], and the outer loop is used to update linearisations of model and observation operators.

For each outer loop l, the inner loop minimises a quadratic objective function to find $\boldsymbol{\delta}\mathbf{x}^{(l)} \in \mathbb{R}^{s(N+1)}$, where $\boldsymbol{\delta}\mathbf{x}^{(l)} = \mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}$. Writing $\boldsymbol{\delta}\mathbf{x} = (\boldsymbol{\delta}\mathbf{x}_0^\top, \boldsymbol{\delta}\mathbf{x}_1^\top, \ldots, \boldsymbol{\delta}\mathbf{x}_N^\top)^\top$, the full non-linear observation operator $\mathcal{H}_i$ (similarly the model operator $\mathcal{M}_i$) is linearised about the current best guess $\mathbf{x}_i^{(l)}$ to obtain the linearised operator $\mathbf{H}_i^{(l)}$ (respectively $\mathbf{M}_i^{(l)}$). The updated initial guess $\boldsymbol{\delta}\mathbf{x}_0^{(l)}$ is propagated forward between observation times by $\mathbf{M}_i^{(l)}$ to obtain $\boldsymbol{\delta}\mathbf{x}_{i+1}^{(l)} = \mathbf{M}_i^{(l)}\boldsymbol{\delta}\mathbf{x}_i^{(l)}$. We note that the time between observations is likely to consist of multiple numerical model time-steps, hence $\mathbf{M}_i^{(l)}$ often corresponds to the composition of many discretised models for a single observation time-step.

Alternatively, the quadratic objective function in the inner loop may be replaced with a saddle point system. Following the notation of [18, Eq. (3.17)], we substitute the linearised objective function with the following saddle point system:

$$\begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{R} & \mathbf{H} \\ \mathbf{L}^\top & \mathbf{H}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta\eta} \\ \boldsymbol{\delta\nu} \\ \boldsymbol{\delta}\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{d} \\ \mathbf{0} \end{pmatrix}. \tag{2.1}$$

In this paper we focus on new preconditioners for the saddle point coefficient matrix:

$$\boldsymbol{\mathcal{A}} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{R} & \mathbf{H} \\ \mathbf{L}^\top & \mathbf{H}^\top & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{(2s+p)(N+1)\times(2s+p)(N+1)}, \tag{2.2}$$

where $\mathbf{D}$, $\mathbf{R}$, $\mathbf{H}$ are the following block diagonal matrices:

$$\begin{aligned} \mathbf{D} &= \texttt{blkdiag}\left(\mathbf{B}, \mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_N\right) \in \mathbb{R}^{(N+1)s\times(N+1)s}, \\ \mathbf{R} &= \texttt{blkdiag}\left(\mathbf{R}_0, \mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N\right) \in \mathbb{R}^{(N+1)p\times(N+1)p}, \\ \mathbf{H} &= \texttt{blkdiag}\left(\mathbf{H}_0^{(l)}, \mathbf{H}_1^{(l)}, \mathbf{H}_2^{(l)}, \ldots, \mathbf{H}_N^{(l)}\right) \in \mathbb{R}^{(N+1)p\times(N+1)s}, \end{aligned}$$

with the ith diagonal block in each case being the relevant matrix for time t_i. Here, the MATLAB-style notation '`blkdiag`' is used to describe a block diagonal matrix in terms of its block diagonal entries.

The matrix $\mathbf{L} \in \mathbb{R}^{(N+1)s\times(N+1)s}$ contains the linearised model information that evolves $\mathbf{x}_0$ between the N observation times, or subwindows, written

$$\mathbf{L} = \begin{pmatrix} \mathbf{I} & & & & \\ -\mathbf{M}_1^{(l)} & \mathbf{I} & & & \\ & -\mathbf{M}_2^{(l)} & \mathbf{I} & & \\ & & \ddots & \ddots & \\ & & & -\mathbf{M}_N^{(l)} & \mathbf{I} \end{pmatrix}, \tag{2.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

where $\mathbf{I}$ denotes the $s \times s$ identity matrix. As we consider preconditioners for the inner loop only, for the remainder of the paper we drop the superscripts that denote the outer loop iteration, and simply use $\mathbf{H}_i$ and $\mathbf{M}_i$.

As the saddle point system is indefinite, methods such as the Conjugate Gradient algorithm cannot be used. Depending on the choice of preconditioner, MINRES [28] or GMRES [35] are examples of viable algorithms for solving linear systems of the form (2.1). We note that one challenge of the saddle point formulation is that monotonic decrease of the objective cost is no longer guaranteed [15]. This can be challenging in operational settings where a very limited number of iterations are performed. In this paper we design new preconditioners and study their numerical performance when iterative methods are permitted to reach convergence. The goal is the design of sufficiently effective and efficient preconditioners to allow convergence of MINRES or GMRES approaches in an operational setting.

Numerical methods for saddle point systems are well-studied in the optimisation literature (see [1] for a comprehensive survey). In order to devise suitable approximations of such systems, as in the forthcoming section, one powerful approach is to approximate the 'leading' $(1,1)$ block of the matrix, along with its Schur complement [25, 27]. For saddle point problems arising from data assimilation, not only does the $(1,1)$ block often have complex structure, but the constraint block contains evolution of the model terms forward/backward in time, leading to a Schur complement which is very difficult to approximate cheaply [13]. Therefore, approximating the constraint block cheaply is an important consideration for a good preconditioner of the matrix (2.2). The work presented here therefore attempts to combine suitable approximations for the $(1,1)$ block, the constraint block, and hence the Schur complement.

2.2. Preconditioners for saddle point systems from data assimilation. We now introduce some preconditioners that have been applied to the saddle point formulation of the data assimilation problem described above. We start by considering two classes of preconditioner: the block diagonal preconditioner and the inexact constraint preconditioner. Although the standard constraint approach [2, 3] would include $\widehat{\mathbf{H}}$, an approximation to $\mathbf{H}$, the use of $\widehat{\mathbf{H}} = \mathbf{0}$ is popular in the data assimilation setting where it is commonly called the 'inexact' constraint preconditioner [13, 15, 14]. In this paper we use the same forms that are considered in [15], which are given by

$$\mathcal{P}_D = \begin{pmatrix} \widehat{\mathbf{D}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{R}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{\mathbf{S}} \end{pmatrix}, \qquad \mathcal{P}_I = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \widehat{\mathbf{L}} \\ \mathbf{0} & \widehat{\mathbf{R}} & \mathbf{0} \\ \widehat{\mathbf{L}}^\top & \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{2.4}$$

where $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{R}}$ are approximations to $\mathbf{D}$ and $\mathbf{R}$ which are easier to apply than the original matrices, and $\widehat{\mathbf{L}}$ is an efficient approximation of $\mathbf{L}$. The exact (negative) Schur complement is given by

$$\mathbf{S} = \mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L} + \mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H},$$

which we may approximate by $\widehat{\mathbf{S}}$. For instance, one may drop the second term and take an approximation $\widehat{\mathbf{L}}$ to $\mathbf{L}$: that is a Schur complement approximation of the form $\widehat{\mathbf{L}}^\top \mathbf{D}^{-1} \widehat{\mathbf{L}}$.

One attraction of the block diagonal preconditioner is that one may guarantee a fixed convergence rate based on the eigenvalues of the preconditioned system $\mathcal{P}_D^{-1}\mathcal{A}$. In Section 3 we present bounds on the eigenvalues of the preconditioned system for the block diagonal preconditioner. The inexact constraint preconditioner has been found to yield improved convergence for both toy and operational-scale data assimilation problems, compared to the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

block diagonal preconditioner for the same choices of $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ [13, 15, 14]. However, it is difficult to develop any theoretical results apart from in the unrealistic setting where $\widehat{\mathbf{L}} = \mathbf{L}$, $\widehat{\mathbf{H}} = \mathbf{H}$, $\widehat{\mathbf{R}} = \mathbf{R}$ [13].

In this paper we consider new preconditioners $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ for the model and observation error terms. We now present choices of $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ that have been studied before. Applying the exact choice of $\mathbf{L}^{-1}$ in a preconditioner can be prohibitively expensive as it requires serial products of model terms and their adjoints [13]. Evaluations of the model acting on a vector can dominate the computational cost of data assimilation algorithms, so reducing the total number of matrix–vector products of the form $\mathbf{M}_i\mathbf{v}$ or $\mathbf{M}_i^\top\mathbf{v}$ is an important consideration when designing new preconditioners. A common choice of preconditioner for $\mathbf{L}$ replaces all of the sub-diagonal terms $\mathbf{M}_i$ with $\mathbf{0}$ [15, 18]. We shall denote this choice of preconditioner as $\mathbf{L}_0$, and note that $\mathbf{L}_0$ is the $(N+1)s \times (N+1)s$ identity matrix, hence so is its inverse.

Another choice of preconditioner that was considered in [13, 15] is given by

$$\mathbf{L}_I = \begin{pmatrix} \mathbf{I} & & & & \\ -\mathbf{I} & \mathbf{I} & & & \\ & -\mathbf{I} & \mathbf{I} & & \\ & & \ddots & \ddots & \\ & & & -\mathbf{I} & \mathbf{I} \end{pmatrix},$$

where the sub-diagonal blocks are replaced with the negative $s \times s$ identity matrix. This was found to perform well experimentally when used in the inexact constraint preconditioner for a two-layer quasi-geostrophic model [13]. However, it does not include any model information, and may not be expected to perform well for all problems. In Section 4 we introduce a new preconditioner that incorporates model information.

Within the outer–inner loop structure of the weak-constraint solver, information from previous outer loops can be used to cheaply update the preconditioners in subsequent loops. This has been found to be computationally beneficial [12]. Such approaches could be combined with the method considered in this paper, and are expected to provide value. However, we consider preconditioners for the first outer loop where there is no additional information, and hence developing computationally feasible stand-alone preconditioners is crucial.

In most prior work the exact observation error covariance matrix has been used in preconditioners. This is due to the fact that the cost of applying $\mathbf{R}^{-1}$ is assumed not to be prohibitive, due to the inherent block diagonal structure of $\mathbf{R}$. However, the rising use of more complex correlation structures (such as inter-channel observation errors, see [45]) means that applying $\mathbf{R}^{-1}$ exactly is not always affordable and may be a computational bottleneck. One option is to apply a computationally cheap approximation of $\mathbf{R}$ within the preconditioner, such as the diagonal of $\mathbf{R}$, however this is typically an inaccurate approximation which can significantly delay convergence. In this paper we consider new classes of preconditioners for the observation error covariance term which we expect to improve convergence in the presence of correlated observation error.

3. Bounds on eigenvalues for the block diagonal preconditioner. For a symmetric positive definite block diagonal preconditioner of the form $\mathcal{P}_D$, it is possible to describe the convergence of the preconditioned MINRES algorithm by analysing the eigenvalues of the preconditioned system $\mathcal{P}_D^{-1}\mathcal{A}$ (see [11, Chapter 4], for instance). In the data assimilation setting $\mathcal{P}_D$ is symmetric positive definite, as all proposed approximations $\widehat{\mathbf{D}}$, $\widehat{\mathbf{R}}$ as well as the Schur complement approximation $\widehat{\mathbf{S}}$ are themselves defined to be symmetric positive definite. We note that for the preconditioned GMRES algorithm spectral information is insufficient to describe convergence (see e.g. [19]). Therefore the theoretical bounds presented in Sections

ETNA
Kent State University and
Johann Radon Institute (RICAM)

170 3–5 are informative strictly for the block diagonal preconditioner, with experiments for the
171 inexact constraint preconditioner $\mathcal{P}_I$ being presented in Section 7 for numerical comparison.

The matrices $\mathcal{A}$ and $\mathcal{P}_D$ defined in (2.2) and (2.4) can be written as

$$\mathcal{A} = \begin{pmatrix} \mathbf{\Phi} & \mathbf{\Psi}^\top \\ \mathbf{\Psi} & \mathbf{0} \end{pmatrix}, \qquad \mathcal{P}_D = \begin{pmatrix} \widehat{\mathbf{\Phi}} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{S}} \end{pmatrix}, \tag{3.1}$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}, \qquad \mathbf{\Psi} = \begin{pmatrix} \mathbf{L}^\top & \mathbf{H}^\top \end{pmatrix},$$

and with $\widehat{\mathbf{\Phi}}$ and $\widehat{\mathbf{S}}$ approximations of $\mathbf{\Phi}$ and the (negative) Schur complement $\mathbf{S} = \mathbf{\Psi}\mathbf{\Phi}^{-1}\mathbf{\Psi}^\top$. We now denote

$$\widetilde{\mathbf{S}} = \mathbf{L}^\top \mathbf{D}^{-1}\mathbf{L}, \qquad \widehat{\mathbf{S}} = \widehat{\mathbf{L}}^\top \mathbf{D}^{-1}\widehat{\mathbf{L}},$$

where $\widehat{\mathbf{D}}$, $\widehat{\mathbf{R}}$, $\widehat{\mathbf{L}}$ are approximations of $\mathbf{D}$, $\mathbf{R}$, $\mathbf{L}$. For the forthcoming theory, we suppose that $\mathbf{D}$, $\widehat{\mathbf{D}}$, $\mathbf{R}$, $\widehat{\mathbf{R}}$, $\mathbf{S}$, $\widetilde{\mathbf{S}}$, $\widehat{\mathbf{S}}$ are symmetric positive definite, with

$$\begin{aligned} &\lambda(\widehat{\mathbf{D}}^{-1}\mathbf{D}) \in [\lambda_{\mathbf{D}}, \Lambda_{\mathbf{D}}], \qquad &&\lambda(\widehat{\mathbf{R}}^{-1}\mathbf{R}) \in [\lambda_{\mathbf{R}}, \Lambda_{\mathbf{R}}], \\ &\lambda(\widetilde{\mathbf{S}}^{-1}\mathbf{S}) \in [\lambda_{\mathbf{S}}, \Lambda_{\mathbf{S}}], \qquad &&\lambda\left((\widehat{\mathbf{L}}^\top\widehat{\mathbf{L}})^{-1}(\mathbf{L}^\top\mathbf{L})\right) \in [\lambda_{\mathbf{L}}, \Lambda_{\mathbf{L}}], \end{aligned}$$

172 where $\lambda(\cdot)$ denotes the eigenvalues of a matrix. We may then prove the following result:

THEOREM 3.1. *With the definitions as stated above, the eigenvalues of $\mathcal{P}_D^{-1}\mathcal{A}$ are real, and satisfy:*

$$\begin{aligned} \lambda(\mathcal{P}_D^{-1}\mathcal{A}) \in &\left[\frac{\lambda_{\mathbf{\Phi}} - \sqrt{\lambda_{\mathbf{\Phi}}^2 + 4\Lambda_{\mathbf{\Phi}}\Lambda_{\mathbf{S}}\Lambda_{\mathbf{L}}\kappa(\mathbf{D})}}{2}, \frac{\Lambda_{\mathbf{\Phi}} - \sqrt{\Lambda_{\mathbf{\Phi}}^2 + \frac{4\lambda_{\mathbf{\Phi}}\lambda_{\mathbf{S}}\lambda_{\mathbf{L}}}{\kappa(\mathbf{D})}}}{2}\right] \\ &\cup [\lambda_{\mathbf{\Phi}}, \Lambda_{\mathbf{\Phi}}] \cup \left[\frac{\lambda_{\mathbf{\Phi}} + \sqrt{\lambda_{\mathbf{\Phi}}^2 + \frac{4\lambda_{\mathbf{\Phi}}\lambda_{\mathbf{S}}\lambda_{\mathbf{L}}}{\kappa(\mathbf{D})}}}{2}, \frac{\Lambda_{\mathbf{\Phi}} + \sqrt{\Lambda_{\mathbf{\Phi}}^2 + 4\Lambda_{\mathbf{\Phi}}\Lambda_{\mathbf{S}}\Lambda_{\mathbf{L}}\kappa(\mathbf{D})}}{2}\right], \end{aligned}$$

173 *where $\lambda_{\mathbf{\Phi}} = \min\{\lambda_{\mathbf{D}}, \lambda_{\mathbf{R}}\}$, $\Lambda_{\mathbf{\Phi}} = \max\{\Lambda_{\mathbf{D}}, \Lambda_{\mathbf{R}}\}$, and $\kappa(\cdot)$ denotes the condition number*
174 *of a matrix.*

Proof. Applying well-known results (see [34, p. 2906], [33, Theorem 4.2.1]), we have

$$\lambda \in \left[\frac{\delta - \sqrt{\delta^2 + 4\Delta\Phi}}{2}, \frac{\Delta - \sqrt{\Delta^2 + 4\delta\phi}}{2}\right] \cup [\delta, \Delta] \cup \left[\frac{\delta + \sqrt{\delta^2 + 4\delta\phi}}{2}, \frac{\Delta + \sqrt{\Delta^2 + 4\Delta\Phi}}{2}\right], \tag{3.2}$$

175 where δ, ϕ denote the minimum eigenvalues of $\widehat{\mathbf{\Phi}}^{-1}\mathbf{\Phi}$, $\widehat{\mathbf{S}}^{-1}\mathbf{S}$ for a general block diagonal
176 saddle point preconditioner (3.1), and Δ, Φ represent the corresponding maximum eigenvalues.

Clearly, for this problem $\delta = \lambda_{\mathbf{\Phi}}$ and $\Delta = \Lambda_{\mathbf{\Phi}}$. It remains to analyse the minimum and maximum eigenvalues of the preconditioned Schur complement, which we note is within the range of the Rayleigh quotient (for $\mathbf{v} \neq \mathbf{0}$):

$$\frac{\mathbf{v}^\top\mathbf{S}\mathbf{v}}{\mathbf{v}^\top\widehat{\mathbf{S}}\mathbf{v}} = \frac{\mathbf{v}^\top\mathbf{S}\mathbf{v}}{\mathbf{v}^\top\widetilde{\mathbf{S}}\mathbf{v}} \cdot \frac{\mathbf{v}^\top\widetilde{\mathbf{S}}\mathbf{v}}{\mathbf{v}^\top\widehat{\mathbf{S}}\mathbf{v}} = \frac{\mathbf{v}^\top\mathbf{S}\mathbf{v}}{\mathbf{v}^\top\widetilde{\mathbf{S}}\mathbf{v}} \cdot \frac{\mathbf{v}^\top\mathbf{L}^\top\mathbf{D}^{-1}\mathbf{L}\mathbf{v}}{\mathbf{v}^\top\mathbf{L}^\top\mathbf{L}\mathbf{v}} \cdot \frac{\mathbf{v}^\top\mathbf{L}^\top\mathbf{L}\mathbf{v}}{\mathbf{v}^\top\widehat{\mathbf{L}}^\top\widehat{\mathbf{L}}\mathbf{v}} \cdot \frac{\mathbf{v}^\top\widehat{\mathbf{L}}^\top\widehat{\mathbf{L}}\mathbf{v}}{\mathbf{v}^\top\widehat{\mathbf{L}}^\top\mathbf{D}^{-1}\widehat{\mathbf{L}}\mathbf{v}}. \tag{3.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Observing that

$$\frac{\mathbf{v}^\top \mathbf{S}\mathbf{v}}{\mathbf{v}^\top \widetilde{\mathbf{S}}\mathbf{v}} \in [\lambda_{\mathbf{S}}, \Lambda_{\mathbf{S}}], \qquad \frac{\mathbf{v}^\top \mathbf{L}^\top \mathbf{D}^{-1}\mathbf{L}\mathbf{v}}{\mathbf{v}^\top \mathbf{L}^\top \mathbf{L}\mathbf{v}} = \frac{\mathbf{y}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{D}\mathbf{y}} \in \left[\frac{1}{\Lambda_{\mathbf{D}}}, \frac{1}{\lambda_{\mathbf{D}}}\right],$$
$$\frac{\mathbf{v}^\top \mathbf{L}^\top \mathbf{L}\mathbf{v}}{\mathbf{v}^\top \widehat{\mathbf{L}}^\top \widehat{\mathbf{L}}\mathbf{v}} \in [\lambda_{\mathbf{L}}, \Lambda_{\mathbf{L}}], \qquad \frac{\mathbf{v}^\top \widehat{\mathbf{L}}^\top \widehat{\mathbf{L}}\mathbf{v}}{\mathbf{v}^\top \widehat{\mathbf{L}}^\top \mathbf{D}^{-1}\widehat{\mathbf{L}}\mathbf{v}} = \frac{\mathbf{z}^\top \mathbf{D}\mathbf{z}}{\mathbf{z}^\top \mathbf{z}} \in [\lambda_{\mathbf{D}}, \Lambda_{\mathbf{D}}],$$

where $\mathbf{y} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{v} \neq \mathbf{0}$, $\mathbf{z} = \mathbf{D}^{-1/2}\widehat{\mathbf{L}}\mathbf{v} \neq \mathbf{0}$, we may write that

$$\frac{\mathbf{v}^\top \mathbf{S}\mathbf{v}}{\mathbf{v}^\top \widehat{\mathbf{S}}\mathbf{v}} \in \left[\frac{\lambda_{\mathbf{S}}\lambda_{\mathbf{L}}\lambda_{\mathbf{D}}}{\Lambda_{\mathbf{D}}}, \frac{\Lambda_{\mathbf{S}}\Lambda_{\mathbf{L}}\Lambda_{\mathbf{D}}}{\lambda_{\mathbf{D}}}\right] = \left[\frac{\lambda_{\mathbf{S}}\lambda_{\mathbf{L}}}{\kappa(\mathbf{D})}, \Lambda_{\mathbf{S}}\Lambda_{\mathbf{L}}\kappa(\mathbf{D})\right].$$

Therefore, it holds that $\phi \geq \frac{\lambda_{\mathbf{S}}\lambda_{\mathbf{L}}}{\kappa(\mathbf{D})}$ and $\Phi \leq \Lambda_{\mathbf{S}}\Lambda_{\mathbf{L}}\kappa(\mathbf{D})$. Substituting the bounds for δ, Δ, ϕ, Φ into (3.2) then gives the result. ∎

REMARK 3.2. Theorem 3.1 is an extension of known results, for example [34, p. 2906] and [33, Theorem 4.2.1], in particular an application of this methodology to saddle point systems arising from weak constraint 4D-Var. We highlight that eigenvalue results of this form are important because they lead to concrete convergence properties of MINRES. As in [11, Theorem 4.14], if $\lambda(\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}}) \in [-\mu_1, -\mu_2] \cup [\mu_3, \mu_4]$, with $\mu_1, \mu_2, \mu_3, \mu_4 > 0$ such that $\mu_1 - \mu_2 = \mu_4 - \mu_3$, then after 2ℓ iterations of MINRES:

$$\|\mathbf{r}^{(2\ell)}\|_{\boldsymbol{\mathcal{P}}_D^{-1}} \leq 2\left(\frac{\sqrt{\mu_1\mu_4} - \sqrt{\mu_2\mu_3}}{\sqrt{\mu_1\mu_4} + \sqrt{\mu_2\mu_3}}\right)^{\ell} \|\mathbf{r}^{(0)}\|_{\boldsymbol{\mathcal{P}}_D^{-1}},$$

where $\mathbf{r}^{(\cdot)}$ denotes the residual vector at a given iteration. Note that both positive or negative intervals for the eigenvalues of $\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}}$ can always be stretched such that $\mu_1 - \mu_2 = \mu_4 - \mu_3$ holds, so this result holds without loss of generality. This allows one to ascertain the convergence behaviour of MINRES when we apply approximations of $\mathbf{L}^\top\mathbf{L}$, $\mathbf{R}$ (as well as $\mathbf{D}$), including those discussed in Sections 4 and 5.

REMARK 3.3. Due to the way the Rayleigh quotient is factored in (3.3), Theorem 3.1 gives a potentially very weak bound when $\mathbf{D}$ is ill-conditioned. We find the main features which affect the quality of the preconditioner, as predicted by the result, are the approximations of $\mathbf{D}$ and $\mathbf{R}$, the effect of dropping the second term of the Schur complement, and the quality of the approximation of $\mathbf{L}^\top\mathbf{L}$ (characterised by the eigenvalues of $\widehat{\mathbf{L}}^{-\top}\mathbf{L}^\top\mathbf{L}\widehat{\mathbf{L}}^{-1}$). The latter quantity is the subject of the forthcoming analysis.

4. Approximations $\widehat{\mathbf{L}}$. Theorem 3.1 suggests that the eigenvalues of the preconditioned system are influenced by the quality of the approximation of $\widehat{\mathbf{L}}^\top\widehat{\mathbf{L}}$ to $\mathbf{L}^\top\mathbf{L}$ in the block diagonal preconditioner. In this section we consider existing and new choices of $\widehat{\mathbf{L}}$ and analyse the eigenvalues and structure of $\widehat{\mathbf{L}}^{-\top}\mathbf{L}^\top\mathbf{L}\widehat{\mathbf{L}}^{-1}$, which is similar to $(\widehat{\mathbf{L}}^\top\widehat{\mathbf{L}})^{-1}(\mathbf{L}^\top\mathbf{L})$. The first choice of $\widehat{\mathbf{L}}$, $\mathbf{L}_0$, has previously been used for saddle point preconditioners for the data assimilation problem [15, 18]. We also propose $\mathbf{L}_M$, a new class of parallelisable preconditioners that depends on a user-defined parameter and incorporates model information.

4.1. A new preconditioner, $\mathbf{L}_M$, and the eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$. We begin by defining our proposed preconditioner, $\mathbf{L}_M$, which incorporates model information explicitly. For a user chosen parameter $1 \leq k \leq N+1$, every kth block sub-diagonal of $\mathbf{L}_M$ (i.e. $\mathbf{M}_k, \mathbf{M}_{2k}, \mathbf{M}_{3k}, \ldots$) is set equal to $\mathbf{0}$. The other entries of $\mathbf{L}_M$ correspond to those of $\mathbf{L}$. Formally we write this as in the definition below.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

DEFINITION 1. *Let $k \in \mathbb{N}$, and let $\mathbf{L}_M = \mathbf{L}_M(k) \in \mathbb{R}^{(N+1)s\times(N+1)s}$ be a block matrix made up of $s \times s$ blocks. For i,j = $1, \dots, N+1$ we define*

$$\text{the } (i,j)\text{th block of } \mathbf{L}_M = \begin{cases} \mathbf{I} & \text{if } i = j, \\ -\mathbf{M}_j & \text{if } i = j+1 \text{ and } j \text{ is not divisible by } k, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Similarly we can write the inverse of $\mathbf{L}_M$ as follows, using straightforward linear algebra:

LEMMA 1. *Let $k \in \mathbb{N}$, and let $\mathbf{L}_M = \mathbf{L}_M(k) \in \mathbb{R}^{(N+1)s\times(N+1)s}$ be a block matrix made up of $s \times s$ blocks. For $i, j = 1, \dots, N+1$ we may evaluate that*

$$\text{the } (i,j)\text{th block of } \mathbf{L}_M^{-1} = \begin{cases} \mathbf{I} & \text{if } i = j, \\ \prod_{m=1}^{i-j} \mathbf{M}_{i-m} & \text{if } 1 \le i-j \le (i-1) \bmod(k), \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

We note that $\mathbf{L}_M^{-1}$ is lower triangular, and both the number of non-zero blocks of $\mathbf{L}_M^{-1}$ and the number of terms in each of the products of $\mathbf{L}_M^{-1}$ are controlled by the parameter k. Note that $\mathbf{L}$, defined in (2.3), satisfies $\mathbf{L} = \mathbf{L}_M(N+1)$ according to this notation. To further justify the effectiveness of this approximation, we now study the eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ theoretically. We begin by stating the structure of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ in terms of the linearised model matrices $\mathbf{M}_i$, using straightforward linear algebra.

LEMMA 2. *We can write $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1} = \mathbf{I} + \mathbf{A}(\mathbf{M})$ where the block entries of $\mathbf{A}(\mathbf{M}) \in \mathbb{R}^{s(N+1)\times s(N+1)}$ are defined as follows. For $n = 1, \dots, \lfloor \frac{N}{k} \rfloor$,*

$$[\mathbf{A}(\mathbf{M})]_{i,j} = \begin{cases} (\prod_{t=i}^{nk} \mathbf{M}_t^\top)(\prod_{q=j}^{nk} \mathbf{M}_{nk-q+j}) & \text{for } (n-1)k+1 \le i,j \le nk, \\ -\prod_{t=j}^{nk} \mathbf{M}_{nk-t+j} & \text{for } i = nk+1, (n-1)k+1 \le j \le nk, \\ -\prod_{t=i}^{nk} \mathbf{M}_t^\top & \text{for } j = nk+1, (n-1)k+1 \le i \le nk, \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

where $[\mathbf{A}(\mathbf{M})]_{i,j}$ denotes the (i,j)th block of $\mathbf{A}(\mathbf{M})$.

We now briefly describe the structure of $\mathbf{A}(\mathbf{M})$. The matrix is made up of $\lfloor \frac{N}{k} \rfloor$ overlapping diagonal blocks, where the size of each block is $(k+1)s \times (k+1)s$. Each block 'overlaps' at the $(nk+1, nk+1)$th block of $\mathbf{A}(\mathbf{M})$, meaning that the maximum number of non-zero blocks in any row or column is given by $2k+1$. We use this structure to demonstrate that our new preconditioner $\mathbf{L}_M$ yields a number of unit eigenvalues for the preconditioned term $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$.

PROPOSITION 1. *Let $\mathbf{L}$ be defined as in (2.3) and $\mathbf{L}_M$ as in Lemma 2. For $2 \le k \le N+1$, $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ has at least rs unit eigenvalues where $r = N+1-2\lfloor \frac{N}{k} \rfloor$.*

Proof. From Lemma 2 we can construct eigenvectors corresponding to zero eigenvalues of $\mathbf{A}(\mathbf{M})$, which will yield unit eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$. Let $\mathbf{e}_t$ define the canonical vector taking unit value in position t and zero elsewhere. Observe that $\mathbf{A}(\mathbf{M})$ is block diagonal matrix with a $(N-k\lfloor \frac{N}{k} \rfloor)s \times (N-k\lfloor \frac{N}{k} \rfloor)s$ zero block in the final position. We can construct $(N-k\lfloor \frac{N}{k} \rfloor)s$ linearly independent eigenvectors corresponding to the zero eigenvalue for this block using $\mathbf{e}_t$ for $t = (k\lfloor \frac{N}{k} \rfloor)s+1, \dots, Ns$.

For each value of n in Lemma 2 we obtain $(k-2)s$ eigenvectors corresponding to a zero eigenvalue, of the form

$$(\mathbf{0}, \dots, \mathbf{0}, \mathbf{v}_t^\top, -(\mathbf{M}_r\mathbf{v}_t)^\top, \mathbf{0}, \dots, \mathbf{0})^\top$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

for $rs+1 \leq t \leq (r+1)s$ and $nk+2 \leq r \leq (n+1)k-1$. Therefore these contribute $(k-2)s\left\lfloor \frac{N}{k} \right\rfloor$ linearly independent eigenvectors corresponding to the zero eigenvalue across the whole matrix. From the first block in the matrix we obtain an additional s linearly independent eigenvectors corresponding to the zero eigenvalue via $(\mathbf{v}_t^\top, -(\mathbf{M}_1\mathbf{v}_t)^\top, \mathbf{0}, \ldots, \mathbf{0})^\top$ for $t = 1, \ldots, s$.

Combining the above reasoning, we obtain rs unit eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ as required, where

$$r = 1 + (k-2)\left\lfloor \frac{N}{k} \right\rfloor + N - k\left\lfloor \frac{N}{k} \right\rfloor = N + 1 - 2\left\lfloor \frac{N}{k} \right\rfloor.$$

□

We see that r does not decrease as k increases, i.e. incorporating information from more timesteps will generally lead to a larger number of unit eigenvalues of the preconditioned model term. Increasing the number of observation times N will broadly lead to an increase in the number of unit eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$, but this behaviour is non-monotonic.

If we introduce assumptions on the spectral radii of the model operator terms, we can obtain explicit bounds on the eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$.

PROPOSITION 2. *If $\|\mathbf{M}_i\mathbf{M}_i^\top\|_2 \leq 1\ \forall i$ then the eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ can be bounded above by $k + 1 + 2\sqrt{k}$.*

Proof. We bound the eigenvalues of $\mathbf{A}(\mathbf{M})$ by splitting the matrix into three sub-matrices $\mathbf{A}(\mathbf{M}) = \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3$, where $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ are symmetric and will be defined explicitly in what follows. As all matrices being considered are symmetric, using [4, Fact 5.12.2] we can bound the maximum eigenvalue of $\mathbf{A}(\mathbf{M})$ above by $\lambda_{\max}(\mathbf{A}(\mathbf{M})) \leq \lambda_{\max}(\mathbf{A}_1) + \lambda_{\max}(\mathbf{A}_2) + \lambda_{\max}(\mathbf{A}_3)$.

Let $\mathbf{A}_1$ be a block diagonal matrix with blocks of size $nk \times nk$, with entries defined by:

$$\text{the } (i,j)\text{th block of } \mathbf{A}_1 = \begin{cases} \text{the } (i,j)\text{th block of } \mathbf{A}(\mathbf{M}) & \text{for } (n-1)k+1 \leq i,j \leq nk, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Each $ks \times ks$ block has rank s, as the first $(k-1)s$ rows are multiples of the final $s \times s$ rows. Substitution yields the eigenvalue problem

$$\left(\mathbf{M}_{nk}^\top \sum_{t=(n-1)k+1}^{nk} \left(\prod_{p=t}^{nk} \mathbf{M}_{nk-p+t}\right)\left(\prod_{q=t}^{nk-1} \mathbf{M}_q^\top\right)\right)\mathbf{v} = \mu\mathbf{v}.$$

We apply [24, Theorem 1.3.20], which states that exchanging the order of matrix multiplication for two compatible matrices has no effect on the non-zero eigenvalues of the product, so we instead consider

$$\left(\sum_{t=(n-1)k+1}^{nk} \left(\prod_{p=t}^{nk} \mathbf{M}_{nk-p+t}\right)\left(\prod_{q=t}^{nk} \mathbf{M}_q^\top\right)\right)\bar{\mathbf{v}} = \mu\bar{\mathbf{v}}.$$

We can separate the contribution of each individual term by applying [4, Fact 5.12.2] to obtain

$$\begin{aligned}\mu &\leq \lambda_{\max}\left(\sum_{t=(n-1)k+1}^{nk} \left(\prod_{p=t}^{nk} \mathbf{M}_{nk-p+t}\right)\left(\prod_{q=t}^{nk} \mathbf{M}_q^\top\right)\right) \\ &\leq \sum_{t=(n-1)k+1}^{nk} \prod_{p=t}^{nk} \|\mathbf{M}_p\mathbf{M}_p^\top\|_2 \leq \sum_{t=(n-1)k+1}^{nk} 1 = k.\end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The maximum eigenvalue of each block is bounded above by k, and hence $\lambda_{\max}(\mathbf{A}_1) \leq k$.

Without loss of generality, assume that $\left\lfloor \frac{N}{k} \right\rfloor$ is odd. For $n = 1, 3, 5, \ldots, 2\left\lfloor \frac{N}{2k} \right\rfloor + 1$ we define $\mathbf{A}_2$ by a block diagonal matrix, with blocks of size $(k+1)s \times (k+1)s$ and entries given by

$$[\mathbf{A}_2]_{i,j} = \begin{cases} \text{the } (i,j)\text{th block of } \mathbf{A}(\mathbf{M}) & \text{for } i = nk+1, (n-1)k+1 \leq j \leq nk, \\ \text{the } (i,j)\text{th block of } \mathbf{A}(\mathbf{M}) & \text{for } j = nk+1, (n-1)k+1 \leq i \leq nk, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

The $(k+1)s \times (k+1)s$ blocks have rank $2s$ with non-zero eigenvalues that solve

$$\left(\sum_{t=(n-1)k+1}^{nk} \left(\prod_{p=t}^{nk} \mathbf{M}_{nk-p+t} \right) \left(\prod_{q=t}^{nk} \mathbf{M}_q^\top \right) \right) \mathbf{v} = \mu^2 \mathbf{v}.$$

By the same argument as above we can bound $\mu^2 \leq k$. Hence $\lambda_{\max}(\mathbf{A}_2) \leq \sqrt{k}$.

For $n = 2, 4, 6, \ldots, 2\left\lfloor \frac{N}{2k} \right\rfloor$, $\mathbf{A}_3$ is a block diagonal matrix with blocks of size $(k+1)s \times (k+1)s$ and entries given by

$$[\mathbf{A}_3]_{i,j} = \begin{cases} \text{the } (i,j)\text{th block of } \mathbf{A}(\mathbf{M}) & \text{for } i = nk+1, (n-1)k+1 \leq j \leq nk-1, \\ \text{the } (i,j)\text{th block of } \mathbf{A}(\mathbf{M}) & \text{for } j = nk+1, (n-1)k+1 \leq i \leq nk-1, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

All blocks have the same structure as the blocks of $\mathbf{A}_2$, and hence have eigenvalues bounded above by $\sqrt{k}$.

The largest eigenvalue of $\mathbf{A}(\mathbf{M})$ is therefore bounded above $\lambda_{\max}(\mathbf{A}(\mathbf{M})) \leq k + 2\sqrt{k}$. By adding the identity matrix, we obtain the upper bound on the eigenvalues in the proposition statement. □

REMARK 4.1. For smaller values of N, $\mathbf{A}_3$ does not enter the working. We can therefore apply a similar argument with $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$ to obtain the tighter bounds $\lambda(\mathbf{L}_M^{-\top} \mathbf{L}^\top \mathbf{L} \mathbf{L}_M^{-1}) \leq 1 + k + \sqrt{k}$ *if* $k \leq N < 2k$, that is $\left\lfloor \frac{N}{k} \right\rfloor = 1$.

REMARK 4.2. A similar approach is not illustrative when examining a lower bound for the eigenvalues, as this would yield negative bounds, whereas all eigenvalues of $\mathbf{L}_M^{-\top} \mathbf{L}^\top \mathbf{L} \mathbf{L}_M^{-1}$ are clearly greater than zero by construction.

5. Approximations $\widehat{\mathbf{R}}_i$. Theorem 3.1 suggests that the eigenvalues of the preconditioned system are also influenced by the quality of the approximation of $\widehat{\mathbf{R}}$ to $\mathbf{R}$ in the block diagonal preconditioner. In this section we consider four choices of $\widehat{\mathbf{R}}_i$, each of which we apply blockwise to $\mathbf{R}_i$. Similarly to the preconditioners for $\mathbf{L}$, we consider an existing choice of preconditioner that is diagonal, and three new choices of preconditioner that include correlation information. We expect the new preconditioners to be beneficial for highly correlated observation error covariance matrices. Correlated observation errors are currently implemented operationally at a number of numerical weather prediction centres for hyperspectral satellite instruments (see e.g. [45, 41]) and Doppler Radar Winds (DRW) (e.g. [36]). Hyperspectral instruments have a block covariance structure with a number of highly correlated off-diagonal entries, and DRW error statistics are spatially correlated. We focus on the observation error covariance matrix in this section as efficient approximations for this term have previously been overlooked. In the following we describe the approaches in terms of $\mathbf{R}$ without loss of generality, as the methods can also be applied to approximate the blocks of $\mathbf{D}$.

REMARK 5.1. We note that $\mathbf{R}$ has a block diagonal structure. Each of the preconditioners in this section is applied blockwise, yielding a block diagonal $\widehat{\mathbf{R}}$ with at least $N+1$ blocks. The two preconditioners presented in Section 5.1 *further* increase the sparsity of $\widehat{\mathbf{R}}$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

5.1. Choices of $\widehat{\mathbf{R}}_i$ which increase sparsity. Many previous studies of the saddle point data assimilation formulation assume that $\mathbf{R}$ is diagonal or easy to invert [15, 12]. For those instruments with uncorrelated or diagonally dominant observation error covariance matrices, it is likely that a diagonal approximation $\widehat{\mathbf{R}}$ will be sufficient. However for more complicated structures, it is unlikely that the exact inverse of $\mathbf{R}$ can be applied efficiently in terms of storage or computation. The approaches we present here are designed to perform well in terms of effectiveness and efficiency for correlation structures that are currently used operationally.

The first two choices of $\widehat{\mathbf{R}}_i$ considered in this section allow for additional sparsification of the observation error component of the preconditioners. The first preconditioner, denoted $\mathbf{R}_{diag}$, takes the diagonal of the original observation error covariance matrix $\mathbf{R}_i$. This is often a first approximation of $\mathbf{R}_i$ or its inverse for simple covariance structures. This choice of $\widehat{\mathbf{R}}_i$ is cheap to apply and extremely sparse, but is expected to perform poorly if there is significant off-diagonal correlation structure.

The second sparsifying choice of $\widehat{\mathbf{R}}_i$, denoted $\mathbf{R}_{block}$ in Algorithm 3 in the Appendix, is designed to exploit existing block structure in $\mathbf{R}_i$. In applications, $\mathbf{R}_i$ itself often has a block structure, with the strength of off-diagonal correlations varying, e.g. for different instruments or measurement types (see e.g. [45]). The idea of $\mathbf{R}_{block}$ is to retain the sub/super-diagonal blocks of $\mathbf{R}_i$ with the largest norm. Neglecting off-diagonal blocks of $\mathbf{R}_i$ with smaller norm ensures that $\mathbf{R}_{block}$ is decoupled into a block diagonal matrix. Let $\mathbf{R}_i \in \mathbb{R}^{p_i \times p_i}$ be a covariance matrix with an associated vector $\texttt{pvec} \in \mathbb{R}^{p_n}$ that specifies the size of 'blocks' such that $\sum_{k=1}^{p_n} \texttt{pvec}_k = p_i$. Algorithm 3 returns a block covariance matrix $\mathbf{R}_{block}$ where only off-diagonal blocks with scaled Frobenius norm larger than a user-defined tolerance $\texttt{tol}$ are retained. In order to maximise computational efficiency only norms of the blocks on the first super-diagonal are computed. If two (or more) adjacent blocks are retained, information from higher level super-diagonals is also included. This does not change the overall block structure of the new preconditioner, but allows for the inclusion of more information from $\mathbf{R}_i$. This will not lead to a large increase in the cost of applying its inverse and we deem that retaining the additional correlation information is likely to be beneficial for the preconditioner.

5.2. Preconditioning methods motivated by reconditioning. The next two choices of $\widehat{\mathbf{R}}_i$ are motivated by reconditioning methods [39]. These are commonly used in data assimilation implementations to mitigate the issues associated with ill-conditioned sample covariance matrices [45]. These methods do not increase the sparsity of $\widehat{\mathbf{R}}_i$ compared to $\mathbf{R}_i$, but can be applied to non-block matrices, such as the spatially varying error covariances used for Doppler Radar Winds. In this application we consider the use of such methods to develop new terms in the preconditioner only.

Algorithm 1: Ridge regression method

Inputs: Matrix $\mathbf{R}_i$, target condition number $\kappa_{\max}$.
Define $\gamma = \frac{\lambda_{\max}(\mathbf{R}_i) - \lambda_{\min}(\mathbf{R}_i)\kappa_{\max}}{\kappa_{\max} - 1}$.
Set $\mathbf{R}_{RR} = \mathbf{R}_i + \gamma \mathbf{I}$.

Algorithms 1 and 2 define parameter-dependent preconditioners. Typically in the reconditioning setting γ and T are selected such that reconditioned matrices have condition number $\kappa_{\max}$. However, in the preconditioning approach we select γ and T directly, with larger parameter values yielding smaller condition numbers of $\widehat{\mathbf{R}}_i$.

Algorithms 1 and 2 can be used to construct preconditioners for $\mathbf{R}_i$ that retain much of the structure of the original matrix, but are better conditioned. Additionally, we can prove how

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 2: Minimum eigenvalue method

Inputs: Matrix $\mathbf{R}_i = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, target condition number $\kappa_{\max}$.
Set $\lambda_{\max}(\mathbf{R}_{ME}) = \lambda_{\max}(\mathbf{R}_i)$.
Define $T = \frac{\lambda_{\max}(\mathbf{R}_i)}{\kappa_{\max}} > \lambda_{\min}(\mathbf{R}_i)$.
Set the remaining eigenvalues of $\mathbf{R}_{ME}$ via

$$\lambda_k(\mathbf{R}_{ME}) = \begin{cases} \lambda_k(\mathbf{R}_i) & \text{if } \lambda_k(\mathbf{R}_i) > T, \\ T & \text{if } \lambda_k(\mathbf{R}_i) \leq T. \end{cases}$$

Construct the reconditioned matrix via $\mathbf{R}_{ME} = \mathbf{V}\mathbf{\Lambda}_{ME}\mathbf{V}^\top$, where $\mathbf{\Lambda}_{ME}$ is a diagonal matrix with diagonal entries given by $\lambda_k(\mathbf{R}_{ME})$.

the eigenvalues of the preconditioned term that appears in Theorem 3.1 change as we vary γ and T, respectively. The following results determine the spectra of the preconditioned terms $\mathbf{R}_{RR}^{-1}\mathbf{R}_i$ and $\mathbf{R}_{ME}^{-1}\mathbf{R}_i$ for any choice of parameters γ and T.

PROPOSITION 3. *Let $\lambda_k(\mathbf{R}_i)$ denote the eigenvalues of $\mathbf{R}_i$. The eigenvalues of $\mathbf{R}_{RR}^{-1}\mathbf{R}_i$ are given by*

$$\lambda_k(\mathbf{R}_{RR}^{-1}\mathbf{R}_i) = \frac{\lambda_k(\mathbf{R}_i)}{\lambda_k(\mathbf{R}_i) + \gamma}.$$

Proof. Let $\mathbf{R}_i = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{\Lambda} = \texttt{diag}(\lambda_k)$, be the eigendecomposition of $\mathbf{R}_i$. Then

$$\mathbf{R}_{RR} = \mathbf{V}(\mathbf{\Lambda} + \gamma\mathbf{I})\mathbf{V}^\top$$

and hence

$$\mathbf{R}_{RR}^{-1}\mathbf{R}_i = \mathbf{V}\left(\texttt{diag}\left(\frac{\lambda_k}{\lambda_k + \gamma}\right)\right)\mathbf{V}^\top.$$

□

We note that for any value of γ, $0 < \lambda(\mathbf{R}_{RR}^{-1}\mathbf{R}_i) < 1$. For small values of γ all eigenvalues are closer to 1 and as γ increases, more eigenvalues move towards zero.

PROPOSITION 4. *The eigenvalues of $\mathbf{R}_{ME}^{-1}\mathbf{R}_i$ are given by*

$$\lambda_k(\mathbf{R}_{ME}^{-1}\mathbf{R}_i) = \begin{cases} 1 & \text{if } \lambda_k > T, \\ \frac{\lambda_k}{T} < 1 & \text{if } \lambda_k \leq T. \end{cases}$$

Proof. Let $\mathbf{R}_i = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{\Lambda} = \texttt{diag}(\lambda_k)$, be the eigendecomposition of $\mathbf{R}_i$. Then

$$\mathbf{R}_{ME} = \mathbf{V}\mathbf{\Lambda}_{ME}\mathbf{V}^\top,$$

where $\mathbf{\Lambda}_{ME}$ is a diagonal matrix with diagonal entries given by $\max\{T, \lambda_k(\mathbf{R}_i)\}$. Hence

$$\mathbf{R}_{ME}^{-1}\mathbf{R}_i = \mathbf{V}\left(\texttt{diag}\left(\frac{\lambda_k}{\max\{T, \lambda_k\}}\right)\right)\mathbf{V}^\top,$$

which has eigenvalues corresponding to the expression in the theorem statement. □

ETNA
Kent State University and
Johann Radon Institute (RICAM)

All of the eigenvalues of $\mathbf{R}_{ME}^{-1}\mathbf{R}_i$ lie in the range $(0, 1]$. For small values of T most of the eigenvalues are units, with a few smaller than 1. As T increases a larger number of eigenvalues become strictly smaller than 1.

Therefore both choices of preconditioner maintain the ordering of eigenvalues and yield eigenvalues of the preconditioned matrix lying between 0 and 1. For both approaches increasing the parameter leads to larger differences between the eigenvalues of the preconditioned matrix and 1. Smaller parameter values may yield choices of $\widehat{\mathbf{R}}$ that are themselves ill-conditioned, and hence expensive to evaluate as part of a preconditioner. Therefore, there is a balance to be struck when choosing a parameter value to avoid poor conditioning of either the preconditioner $\widehat{\mathbf{R}}$ or the preconditioned term $\widehat{\mathbf{R}}^{-1}\mathbf{R}$. A natural question is therefore how to select appropriate parameter values, and how to implement this automatically. Heuristics for automated parameter selection are discussed in Section 6 for our numerical experiments, but it is likely that some initial investigation would be necessary to identify suitable methods for specific problems of interest. In Section 7 we compare the performance of the four approximations $\mathbf{R}_{diag}$, $\mathbf{R}_{block}$, $\mathbf{R}_{RR}$, and $\mathbf{R}_{ME}$, applied block-wise to each block $\mathbf{R}_i$ of $\mathbf{R}$.

6. Numerical framework.

In this section we introduce the numerical framework for the experiments presented in Section 7. We begin in Section 6.1 by defining the parameters for the data assimilation problem. The same data assimilation framework is used for all experiments. In Section 6.2 we discuss implementation aspects relating to the preconditioners. We note that all results are computed using MATLAB version 2019b on a machine an a 1.8GHz Intel Intel quad-core i7 processor with 15GB RAM on an Ubuntu 20.04.2 LTS operating system.

6.1. Data assimilation parameters.

We now describe the data assimilation problem that is studied in Section 7. The size of the state space s is determined by the spatial discretisation $s = \frac{1}{\Delta x}$. For each choice of s we fix the observation operator $\mathbf{H}_i \in \mathbb{R}^{p\times s}$ to be the same for all observation times, i. We choose $p = \frac{s}{2}$, and observations of alternate state variables are smoothed equally over 5 adjacent state variables, with entries either being 0 or $\frac{1}{5}$. The full observation operator $\mathbf{H}$ is then assembled by taking $\mathbf{H} = \mathbf{I}_{N+1} \otimes \mathbf{H}_i$, where $\otimes$ denotes the Kronecker product.

We assume that the model error $\mathbf{Q}_i$ is the same at each observation time, i.e. $\mathbf{Q}_i \equiv \mathbf{Q}$ for $i = 1, \ldots, N$. Although this is a simplified choice of model error, we note that treating $\mathbf{Q}_i$ is not the focus of this work. More complicated formulations could be taken into account for operational problems, and the preconditioning approaches for $\mathbf{R}$ discussed in Section 5 can also be applied to the blocks of $\mathbf{Q}$. Both $\mathbf{B}$ and $\mathbf{Q}$ are created using the same routine, based on a SOAR correlation matrix [46]. This routine constructs spatial local correlations whilst ensuring that the matrix has high sparsity. Both $\mathbf{B}$ and $\mathbf{Q}$ are $s \times s$ circulant matrices fully defined by a single row. The number of non-zero entries in each row is fixed irrespective of the value of s. The non-zero entries are computed using a modified SOAR function following the procedure in [22]:

$$c_i = \sigma\left(1 + \frac{2|\sin\left(\frac{i\theta}{2}\right)|}{L}\exp\left(-\frac{2|\sin(\frac{i\theta}{2})|}{L}\right)\right), \qquad \theta = \frac{\pi}{\texttt{maxval}}, \tag{6.1}$$

where L is the correlation lengthscale (0.6 for $\mathbf{B}$, 0.5 for $\mathbf{Q}$), `maxval` determines the number of non-zero entries (100 for $\mathbf{B}$, 120 for $\mathbf{Q}$), and σ is the amplitude of the correlation function (0.4 for $\mathbf{B}$, 0.2 for $\mathbf{Q}$). To ensure positive definiteness of the function, if the smallest eigenvalue of the full matrix is negative a constant $\delta = |\lambda_{\min}(\mathbf{B})| + \psi$ is added to the diagonal, where ψ is a random number in $[0, 0.5]$. We then assemble $\mathbf{D}$ by taking the first block diagonal entry to be $\mathbf{B}$ and the remaining block diagonal entries to be N copies of $\mathbf{Q}$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Experiment	Model	$\mathbf{D}$	$\mathbf{R}_i$	$\mathbf{H}$	p/s
A	Lorenz 96	Modified SOAR	Block	Direct observations	0.5
B	Heat equation	Modified SOAR	Block	Direct observations	0.25

TABLE 6.1

Summary of experimental design for Section 7. The modified SOAR function is described in (6.1), the block method to construct $\mathbf{R}_i$ is described in Section 6.1, as is the direct observation structure of $\mathbf{H}$. The final column shows the ratio of observations to state variables at each observation time.

The spatial structure for $\mathbf{B}$ and $\mathbf{Q}$ introduced above is similar to numerical frameworks that have been considered previously for weak-constraint data assimilation experiments [8, 10]. However, we note that for realistic data assimilation systems the true structure of $\mathbf{Q}_i$ is not well known. Improved understanding of model error covariance matrices, and their estimation in preconditioners, is of research interest, but will not be considered here. For our experiments we apply the ridge regression preconditioner to $\mathbf{B}$ and $\mathbf{Q}$ given by Algorithm 1, for the inflation parameter $\gamma = 0.01$ to obtain $\mathbf{B}_{RR}$ and $\mathbf{Q}_{RR}$. We apply $\widehat{\mathbf{D}}^{-1}$ using the incomplete Cholesky factors of $\mathbf{B}_{RR}$ and $\mathbf{Q}_{RR}$, respectively, computed with the MATLAB function `ichol` with zero-fill, i.e. using the same sparsity structure as $\mathbf{B}$ and $\mathbf{Q}$.

Many experiments that account for correlated observation error covariance matrices use spatial correlations and circulant matrix structures (similar to those we are using for $\mathbf{B}, \mathbf{Q}$). However, in NWP error correlations often arise from hyperspectral satellite-based instruments which have interchannel uncertainty structures (see for instance [45, 37]), which appear as block structures within a matrix. Therefore for these experiments we construct a matrix with block structure, designed to replicate many of the properties of realistic interchannel error correlations.

To construct $\mathbf{R}_i \in \mathbb{R}^{p\times p}$ we define two vectors: $\texttt{pvec} \in \mathbb{R}^{\texttt{plen}}$ such that $p = \sum_{k=1}^{\texttt{plen}} \texttt{pvec}_k$ which gives the size of the blocks, and $\texttt{pcorr} \in \mathbb{R}^{\texttt{plen}(\texttt{plen}-1)/2}$ which gives the a multiplication factor for each off-diagonal block (note if $\texttt{pcorr}_k = 0$ the corresponding off-diagonal block is uncorrelated). Diagonal blocks are correlated and constructed as the Hadamard product of a sparse random matrix and a sparse SOAR matrix (using the same approach as for $\mathbf{B}$ and $\mathbf{Q}$ above). Off-diagonal blocks are sparse random matrices with entries in $(0, \texttt{pcorr}_k)$. The matrix $\mathbf{R}_i$ is assembled by adding the diagonal blocks and upper half of the matrix and then symmetrising. This also increases the weight of the diagonal blocks. A similar approach to that used for $\mathbf{B}$ and $\mathbf{Q}$ is applied to guarantee that $\mathbf{R}_i$ is positive definite, where if the minimum eigenvalue of $\mathbf{R}_i$ is less than 1 it is increased to a small positive value. This value is fixed at 0.41, in order to control the conditioning of $\mathbf{R}_i$, and ensure that the condition numbers of $\mathbf{D}$ and $\mathbf{R}$ are comparable. We assemble $\mathbf{R}$ by taking $\mathbf{R} = \mathbf{I}_{N+1} \otimes \mathbf{R}_i$, where $\otimes$ denotes the Kronecker product and the choice of $\mathbf{R}_i \in \mathbf{R}^{p\times p}$ is fixed for a given data assimilation problem. This method of construction ensures we have adequate sparsity for high-dimensional experiments, that $\mathbf{R}_i$ is positive definite with significant correlations, and that it is well-conditioned. In practice $\mathbf{R}$ may be very ill-conditioned compared to $\mathbf{D}$, in which case we expect selecting a good choice of $\widehat{\mathbf{R}}$ would be even more vital to ensure fast convergence. The combination of model and data assimilation parameters used in our experiments is summarised in Table 6.1.

Finally, we wish to apply the reconditioning-inspired preconditioner to $\mathbf{R}_i$ in an online way. One way to do this is to use the small eigenvalues of $\mathbf{R}_i$ to select γ and T. For the ridge regression approach, we set $\gamma = \lambda_{\min}(\mathbf{R}_i)$. For the minimum eigenvalue method we compute the smallest two eigenvalue–eigenvector pairs, and set the threshold equal to the second smallest, meaning only a single eigenvalue is changed. The small eigenvalues are computed using `eigs(R,1,'sr')` in MATLAB. For both approaches we use information from $\mathbf{R}_i$, but computing a small number of eigenvalues ensures this is computationally efficient.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

	$\mathbf{R}_i$		$\mathbf{R}_{diag}^{-1}\mathbf{R}_i$		$\mathbf{R}_{block}^{-1}\mathbf{R}_i$		$\mathbf{R}_{RR}^{-1}\mathbf{R}_i$		$\mathbf{R}_{ME}^{-1}\mathbf{R}_i$
p	$\lambda_{\min}$	$\lambda_{\max}$	$\lambda_{\min}$	$\lambda_{\max}$	$\lambda_{\min}$	$\lambda_{\max}$	$\lambda_{\min}$	$\lambda_{\max}$	$\lambda_{\min}$
125	0.4100	108.63	0.0131	2.7466	0.6779	1.3221	0.2908	1.0000	0.2602
250	0.4100	200.55	0.0077	3.3466	0.7886	1.2114	0.2908	1.0000	0.6021
500	0.4100	312.05	0.0061	4.4236	0.8889	1.1110	0.2908	1.0000	0.1129
1250	0.4100	540.99	0.0044	5.5440	0.9335	1.0665	0.2908	1.0000	0.1077
2500	0.4100	650.67	0.0055	6.0254	0.7556	1.2444	0.2908	1.0000	0.1302
5000	0.4100	706.94	0.0057	6.8106	0.6058	1.3941	0.2908	1.0000	0.1725

TABLE 6.2

Minimum and maximum eigenvalues of preconditioned $\mathbf{R}_i$ for different choices of $\widehat{\mathbf{R}}_i$ computed using the `eigs` *function in* MATLAB.

Finally, for all correlated choices of $\widehat{\mathbf{R}}_i$ we apply $\widehat{\mathbf{R}}_i^{-1}$ using the incomplete Cholesky factors computed with the MATLAB function `ichol` with the same sparsity structure as $\mathbf{R}_i$. For $\mathbf{R}_{ME}$ this means the reconditioning method is applied as a low-rank update to the Cholesky factors via the Woodbury identity as this is more efficient in terms of storage.

Table 6.2 shows the extreme eigenvalues of $\widehat{\mathbf{R}}_i^{-1}\mathbf{R}_i$ for each of the preconditioners discussed in Section 5. We recall that the maximum eigenvalue of $\mathbf{R}_{ME}^{-1}\mathbf{R}_i$ is 1 by definition. We fix the smallest eigenvalue of $\mathbf{R}_i$, $\lambda_{\min}(\mathbf{R}_i) = 0.41$, to ensure that $\mathbf{R}_i$ is well-conditioned. We see that the maximum eigenvalue of $\mathbf{R}_i$ increases with p. Eigenvalues are more extreme for $\mathbf{R}_{diag}^{-1}\mathbf{R}_i$ than for any other preconditioned matrix. Including correlation information is beneficial in terms of the extreme eigenvalues. As the parameter choice for $\mathbf{R}_{RR}$ only depends on the smallest eigenvalue of $\mathbf{R}_i$ (which is fixed), the minimum and maximum eigenvalues of $\mathbf{R}_{RR}^{-1}\mathbf{R}_i$ do not change with increasing p. The block approach clusters both minimum and maximum eigenvalues of $\mathbf{R}_{block}^{-1}\mathbf{R}_i$ either side of 1, whereas the reconditioning approaches lead to a maximum eigenvalue of the preconditioned matrix which is very close or equal to 1. For the approach used here, where T is set to the second smallest eigenvalue of $\mathbf{R}_i$, all eigenvalues bar the smallest of $\mathbf{R}_{ME}^{-1}\mathbf{R}_i$ are equal to 1.

6.2. Aspects of numerical linear algebra implementation. In this section we briefly discuss some of the numerical linear algebra aspects of our implementation to run the numerical experiments in Section 7.

We take advantage of the specific matrix structures in (2.2), and store only the non-zero blocks of the matrix, i.e. $\mathbf{R}_i \in \mathbf{R}^{p\times p}$, $\mathbf{B}, \mathbf{Q}_i \in \mathbf{R}^{s\times s}$ and $\mathbf{H}_i \in \mathbf{R}^{p\times s}$. Each of these (relatively) small matrices is stored as a sparse matrix. We also precompute and store $\widehat{\mathbf{R}}_i \in \mathbf{R}^{p\times p}$ prior to the iteration of the Krylov subspace method.

We compute the matrix–vector products $\boldsymbol{\mathcal{A}}\mathbf{v}$, and the preconditioner solves $\boldsymbol{\mathcal{P}}_D^{-1}\mathbf{v}$ and $\boldsymbol{\mathcal{P}}_I^{-1}\mathbf{v}$, via the action of a matrix on a vector rather than building the full matrices. We now describe this process briefly. Note first that

$$\boldsymbol{\mathcal{A}}\mathbf{v} = \boldsymbol{\mathcal{A}}\begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\\ \mathbf{v}_3\end{pmatrix} := \begin{pmatrix}\mathbf{c}_1\\ \mathbf{c}_2\\ \mathbf{c}_3\end{pmatrix} = \begin{pmatrix}\mathbf{D}\mathbf{v}_1 + \mathbf{L}\mathbf{v}_3\\ \mathbf{R}\mathbf{v}_2 + \mathbf{H}\mathbf{v}_3\\ \mathbf{L}^\top\mathbf{v}_1 + \mathbf{H}^\top\mathbf{v}_2\end{pmatrix}.$$

We compute $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ separately by looping over the $N+1$ blocks of $\mathbf{D}$, $\mathbf{R}$, and $\mathbf{H}$. The action of $\mathbf{L}$ and $\mathbf{L}^{-1}$ on a vector is applied via a function. We note that for each evaluation of the matrix–vector product we require one evaluation each of $\mathbf{L}^\top$ and $\mathbf{L}$.

To apply the inverse of the block diagonal preconditioner to a vector, we have that

$$\boldsymbol{\mathcal{P}}_D^{-1}\mathbf{v} = \begin{pmatrix}\widehat{\mathbf{D}}^{-1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{R}}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \widehat{\mathbf{S}}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{v}_1\\ \mathbf{v}_2\\ \mathbf{v}_3\end{pmatrix} = \begin{pmatrix}\widehat{\mathbf{D}}^{-1}\mathbf{v}_1\\ \widehat{\mathbf{R}}^{-1}\mathbf{v}_2\\ \widehat{\mathbf{L}}^{-1}\mathbf{D}\widehat{\mathbf{L}}^{-\top}\mathbf{v}_3\end{pmatrix}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We see that each application of $\mathcal{P}_D^{-1}$ to a vector requires one evaluation each with $\widehat{\mathbf{D}}^{-1}, \widehat{\mathbf{R}}^{-1}, \widehat{\mathbf{L}}^{-1}$, and $\widehat{\mathbf{L}}^{-\top}$. To apply $\widehat{\mathbf{D}}^{-1}$ and $\widehat{\mathbf{R}}^{-1}$ we loop over the $N+1$ blocks of the matrices, and for $\widehat{\mathbf{L}}^{-1}$ and $\widehat{\mathbf{L}}^{-\top}$ we loop over the k blocks of the relevant sub-matrices. Now,

$$\mathcal{P}_I^{-1}\mathbf{v} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \widehat{\mathbf{L}}^{-\top} \\ \mathbf{0} & \widehat{\mathbf{R}}^{-1} & \mathbf{0} \\ \widehat{\mathbf{L}}^{-1} & \mathbf{0} & -\widehat{\mathbf{S}}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{pmatrix} = \begin{pmatrix} \overbrace{\widehat{\mathbf{L}}^{-\top}\mathbf{v}_3}^{\mathbf{c}_1} \\ \widehat{\mathbf{R}}^{-1}\mathbf{v}_2 \\ \widehat{\mathbf{L}}^{-1}(\mathbf{v}_1 - \mathbf{D}\mathbf{c}_1) \end{pmatrix}.$$

By computing the first component of $\mathcal{P}_I^{-1}\mathbf{v}$, $\mathbf{c}_1$, prior to the final component, we can apply the inexact constraint preconditioner with the same dominant costs-per-iteration as an application of the block diagonal preconditioner. This saves the cost of a second application of $\widehat{\mathbf{L}}^{-\top}$. An additional advantage of the inexact constraint preconditioner is that it does not require the application of $\widehat{\mathbf{D}}^{-1}$ to a vector, and hence may result in significant computational savings for challenging choices of $\mathbf{D}$.

We note that all of the experiments in Section 7 run the Krylov subspace methods to convergence to a relative tolerance of 10^{-6}. This is in contrast to most operational data assimilation implementations, where a small number of fixed iterations are applied. Due to the inclusion of the Lagrange multipliers in the residual of the saddle point formulation, we no longer have monotonic decrease of the objective cost. Running our experiments to convergence ensures a fair comparison between each preconditioner. We leave the investigation of the non-monotonicity for each preconditioner to future work.

7. Numerical experiments. In this section we present numerical experiments for the two problems described in Section 6. We use the MINRES implementation of [29] for the block diagonal preconditioner, with a residual-based convergence criterion in the two norm. For the inexact constraint preconditioner we use the GMRES implementation of [20] with no restarts, and a convergence criterion given by the relative residual in the two norm. We use a tolerance of 10^{-6} for both problems of interest.

We note that all of our experiments converge within the maximum number of iterations we allow (1000).

7.1. Lorenz 96 model. The main problem of interest concerns the Lorenz 96 model [26], a non-linear problem that is often considered as a test problem for data assimilation applications (see for example [8, 18] for use within the saddle point formulation for data assimilation). The Lorenz 96 model consists of s coupled ordinary differential equations which are discrete in space and continuous in time. Consider s equally spaced points on the unit line, i.e. $\Delta x = \frac{1}{s}$. For $i = 1, \dots, s$,

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + 8,$$

where we have periodic boundary conditions (so $x_{-1} = x_{s-1}$, $x_0 = x_s$, $x_{s+1} = x_1$). The choice of forcing constant $F = 8$ induces chaotic behaviour, and is typical for data assimilation applications. We use the numerical implementation of [10], where the model is integrated in time with a fourth-order Runge–Kutta scheme. For all experiments we consider $N = 15$ subwindows. We consider $\Delta t = 10^{-4}$ for all the experiments, although similar results were obtained for other values of Δt and are not presented here.

As we are interested in assessing the performance of preconditioners within the linearised inner loops, we consider a single outer loop of the weak constraint formulation. The Lorenz 96 example can be considered as a study of how well the proposed preconditioners perform

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$\widehat{\mathbf{L}}$	$\widehat{\mathbf{R}}$	$\lambda_{\min}(\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}})$	$\max(\lambda(\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}})<0)$	$\min(\lambda(\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}})>0)$	$\lambda_{\max}(\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}})$
$\mathbf{L}_M$	$\mathbf{R}_{block}$	-15.3994	-0.1233	0.7886	16.4061
$\mathbf{L}_M$	$\mathbf{R}_{RR}$	-14.2002	-0.0574	0.2908	14.8536
$\mathbf{L}_M$	$\mathbf{R}$	-13.8536	-0.1773	1.0000	14.8536
$\mathbf{L}_0$	$\mathbf{R}$	-12.2360	-0.0551	1.0000	13.2360
$\mathbf{L}_M$	$\mathbf{R}_{block}$	-3.5516	-0.1773	0.8295	4.5535
$\mathbf{L}_M$	$\mathbf{R}_{RR}$	-2.5351	-0.1773	0.3878	3.2438
$\mathbf{L}_M$	$\mathbf{R}$	-3.6119	-0.1773	1.0000	4.6119
$\mathbf{L}_0$	$\mathbf{R}$	-2.0415	-0.0551	1.0000	3.0415

TABLE 7.1

Experiment A: (Top:) Bounds on negative and positive eigenvalues from Theorem 3.1 for $s = 500$, $N = 5$, $k = 3$ with parameters from Experiment A in Table 6.1; (Bottom:) Extreme negative and positive eigenvalues for this problem computed using the `eigs` *function in* MATLAB.

k	$\mathbf{R}_i$	$\mathbf{D}_i$	$\widehat{\mathbf{D}}_i^{-1}$	$\mathbf{M}_i/\mathbf{M}_i^\top$	$\mathbf{R}_i$	$\mathbf{R}_{block}^{-1}$	$\mathbf{D}_i$	$\widehat{\mathbf{D}}_i^{-1}$	$\mathbf{M}_i/\mathbf{M}_i^\top$
1	22496	44992	22496	42180	10704	10704	21408	10704	20070
2	18368	36736	18368	34440	8240	8240	16480	8240	15450
3	16688	33376	16688	47978	7536	7536	15072	7536	21666
4	13168	26336	13168	41150	6624	6624	13248	6624	20700
7	11776	23552	11776	41216	6080	6080	12160	6080	21280
10	9520	19040	9520	34510	4816	4816	9632	4816	17458
16	3520	7040	3520	12760	1376	1376	2752	1376	4988

TABLE 7.2

Experiment A: Total number of matrix–vector products with component matrices for $\boldsymbol{\mathcal{P}}_D$ for increasing k for $\mathbf{R}_{diag}$ (left) and $\mathbf{R}_{block}$ (right).

in a realistic setting where the linearised model operators $\mathbf{M}_i$ differ in each subwindow. We note that the setting of of Proposition 1 holds, and hence using $\mathbf{L}_M$ guarantees that $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ possesses a number of unit eigenvalues depending on the choice of k. However, we cannot bound the maximum eigenvalue of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ using the theory of Section 4: the assumptions of Proposition 2 are not satisfied as $\lambda_{\max}(\mathbf{M}_{15}^\top\mathbf{M}_{15}) > 1$ for all choices of Δx that were studied.

Table 7.1 shows the values of the bounds from Theorem 3.1 (top four rows) and the computed eigenvalues (bottom four rows) when using $\boldsymbol{\mathcal{P}}_D$ with $\mathbf{L}_M(3)$, and a number of choices for $\widehat{\mathbf{R}}$ introduced in Section 5. We note that these experiments consider $\mathbf{D} = \mathbf{I}$, as the condition number of $\mathbf{D}$ appears in all terms in the bound. In the more realistic case where $\mathbf{B}, \mathbf{Q}_i \neq \mathbf{I}$ we expect the bounds given by Theorem 3.1 to be much weaker. We note that even with this choice of $\mathbf{D}$ the bounds provide pessimistic estimates of the eigenvalues. However, the qualitative behaviour of the bounds is similar to that of the computed eigenvalues for different choices of $\widehat{\mathbf{R}}$ and $\mathbf{L}_M$. The eigenvalues of smallest absolute value are mainly affected by the choice of $\mathbf{L}_M$ (for the negative eigenvalue with largest magnitude) or $\widehat{\mathbf{R}}$ (for the smallest positive eigenvalue). This is not true for the bounds in the case of the negative eigenvalue with smallest magnitude. The computed eigenvalues of largest magnitude and the corresponding bounds are affected by changes to both $\mathbf{L}_M$ and $\widehat{\mathbf{R}}$, but these changes are rather small in all cases. As the largest magnitude computed eigenvalues are small for this problem (all less than five), the improvements to the small magnitude eigenvalues with $\mathbf{L}_M$ and correlated choices of $\widehat{\mathbf{R}}$ are likely to have the most significant influence on the convergence of the iterative methods.

We now study the performance of the proposed choices of preconditioners for the Lorenz 96 problem (Experiment A in Table 6.1). The left panel of Figure 7.1 shows how the number of iterations required for convergence changes with the choice of $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ within $\boldsymbol{\mathcal{P}}_D$ as k

ETNA
Kent State University and
Johann Radon Institute (RICAM)

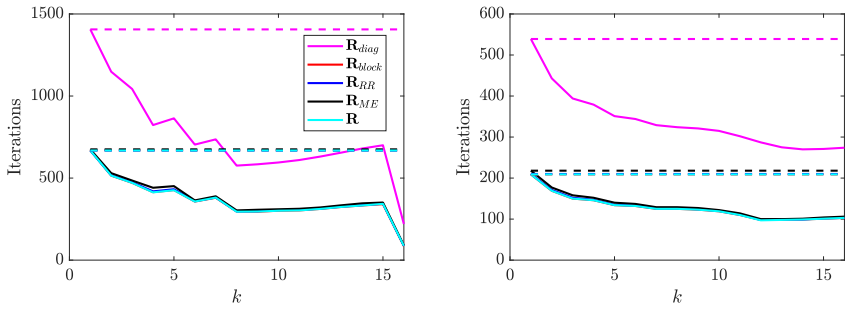

FIG. 7.1. *Experiment A: Number of iterations required for convergence of MINRES for different choices of* $\widehat{\mathbf{R}}$ *within the block diagonal preconditioner* $\boldsymbol{\mathcal{P}}_D$ *(left) and the the inexact constraint preconditioner* $\boldsymbol{\mathcal{P}}_I$ *(right) for the Lorenz 96 problem, with increasing* k*. The dimension of the problem is given by* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{100,000\times 100,000}$*. The dashed line shows the results for* $k = 1$ *i.e.* $\mathbf{L}_M = \mathbf{I} = \mathbf{L}_0$*. We note that* $k = 16$ *corresponds to* $\mathbf{L}_M \equiv \mathbf{L}$*. Legend entries correspond to different choices of* $\widehat{\mathbf{R}}$ *as described in Section 5.*

k	$\mathbf{R}_i$	$\mathbf{D}_i$	$\mathbf{M}_i/\mathbf{M}_i^\top$	$\mathbf{R}_i$	$\mathbf{R}_{block}^{-1}$	$\mathbf{D}_i$	$\mathbf{M}_i/\mathbf{M}_i^\top$
1	8624	17248	16170	3344	3344	6688	6270
2	7088	14176	13290	2704	2704	5408	5070
3	6304	12608	18124	2400	2400	4800	6900
4	6064	12128	18950	2336	2336	4672	7300
7	5264	10528	18424	2000	2000	4000	7000
10	5040	10080	18270	1904	1904	3808	6902
16	4384	8768	15892	1648	1648	3296	5974

TABLE 7.3

Experiment A: Total number of matrix–vector products with component matrices for $\boldsymbol{\mathcal{P}}_I$ *for increasing* k *for* $\mathbf{R}_{diag}$ *(left) and* $\mathbf{R}_{block}$ *(right).*

increases. Including some model information in $\widehat{\mathbf{L}}$ leads to a reduction in iterations compared to $k = 1$ for all choices of $\widehat{\mathbf{R}}$. For $k < 5$ including more model information leads to faster convergence. However, for $k \geq 5$ the change in iterations is non monotonic. Including correlation information in $\widehat{\mathbf{R}}$ results in large improvements to convergence. We note that the lines corresponding to the correlated choices of $\widehat{\mathbf{R}}$ ($\mathbf{R}_{block}$, $\mathbf{R}_{RR}$, and $\mathbf{R}_{ME}$) lie almost directly on top of the line for $\mathbf{R}$. Indeed, if using $\mathbf{R}_{diag}$ we require $k \geq 8$ to obtain fewer iterations than using $\mathbf{L}_0$ with an improved choice of $\widehat{\mathbf{R}}$. There is very little difference in performance between the correlated choices of $\widehat{\mathbf{R}}$. For all choices of $\widehat{\mathbf{R}}$ the smallest number of iterations occurs when using $k = N + 1$, i.e. the exact choice of $\mathbf{L}$. Table 7.2 shows the number of matrix–vector products required to reach convergence for the block diagonal preconditioner using $\mathbf{R}_{diag}$ and $\mathbf{R}_{block}$. Results using $\mathbf{R}_{RR}$ and $\mathbf{R}_{ME}$ are similar to those with $\mathbf{R}_{block}$. The number of matrix–vector products with $\mathbf{D}_i$, $\widehat{\mathbf{D}}_i^{-1}$ and $\mathbf{M}_i$ is reduced when using $\mathbf{R}_{block}$ compared to $\mathbf{R}_{diag}$. For some choices of k the total number of evaluations with $\mathbf{R}_i$ and $\mathbf{R}_{block}^{-1}$ is slightly larger than in the $\mathbf{R}_{diag}$ case. Increasing k broadly decreases the number of matrix–vector products with the error covariance matrices and their inverses. For some choices of $k > 1$ more evaluations of $\mathbf{M}_i$ and $\mathbf{M}_i^\top$ are required than when using $\mathbf{L}_0$. However, this increase is small compared to the decrease in the other components.

The right panel of Figure 7.1 shows how the number of iterations required for convergence changes with the choice of $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ within $\boldsymbol{\mathcal{P}}_I$ as k increases. We see a clear benefit of including model information in terms of a reduction in iterations. Increasing k leads to a reduction in the number of iterations required for convergence when using $\mathbf{L}_M$ for all choices

ETNA
Kent State University and
Johann Radon Institute (RICAM)

	$\mathbf{R}_{block}$	$\mathbf{R}_{RR}$	$\mathbf{R}$	$\mathbf{R}_{block}$	$\mathbf{R}_{RR}$	$\mathbf{R}$
$\mathbf{L}_0$	759	822	822	359	275	275
$\mathbf{L}_M$, $k=3$	433	466	467	244	205	205
$\mathbf{L}_M$, $k=4$	348	335	336	228	200	200
$\mathbf{L}_M$, $k=5$	367	354	355	206	182	182

TABLE 7.4

Experiment A: Number of iterations required for convergence of MINRES with the block diagonal preconditioner $\mathcal{P}_D$ (left) and $\mathcal{P}_I$ (right) applied to the Lorenz 96 problem, using $\mathbf{R}_{block}$, $\mathbf{R}_{RR}$, $\mathbf{R}$ in combination with $\mathbf{L}_0$, $\mathbf{L}_M$ ($k=3,4,5$). Here, $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{1,600,000\times 1,600,000}$.

of $\widehat{\mathbf{R}}$, unlike when using $\mathcal{P}_D$. The benefit of using an improved estimate of $\widehat{\mathbf{R}}$ is even more stark, with any choice of correlated $\widehat{\mathbf{R}}$ and $\mathbf{L}_0$ leading to fewer iterations than $\mathbf{R}_{diag}$ even when using $\widehat{\mathbf{L}} \equiv \mathbf{L}$. Again, the lines for the correlated choices of $\widehat{\mathbf{R}}$ ($\mathbf{R}_{block}$, $\mathbf{R}_{RR}$, and $\mathbf{R}_{ME}$) lie almost directly on top of the line for $\mathbf{R}$. Table 7.3 shows the total number of matrix–vector products required to reach convergence for the inexact constraint preconditioner for $\mathbf{R}_{diag}$ and $\mathbf{R}_{block}$. Results using $\mathbf{R}_{RR}$ and $\mathbf{R}_{ME}$ are similar to those with $\mathbf{R}_{block}$. In this case, using $\mathbf{R}_{block}$ leads to a large reduction in the number of matrix–vector products for all components. Increasing k can lead to increases in the number of model matrix–vector evaluations, but leads to decreases in the total number of matrix–vector products.

Table 7.4 shows the performance of the block diagonal preconditioner and inexact constraint preconditioner, respectively, for a higher-dimensional problem when using $\mathbf{R}_{block}$, $\mathbf{R}_{RR}$, and $\mathbf{R}$ itself to approximate $\mathbf{R}$ within the preconditioner. Similarly to the smaller dimensional problem considered in Figure 7.1 using $\mathbf{L}_M$ leads to improved convergence in terms of iterations compared to $\mathbf{L}_0$. The different choices of $\widehat{\mathbf{R}}$ lead to comparable iteration numbers and we recall that $\mathbf{R}_{block}$ has additional sparsity structure. Increasing k leads to a slight reduction in the number of iterations, but increases the computational cost of each iteration. For this problem, choosing $k=3$ or 4 allows decreased iteration counts compared to $\mathbf{L}_0$, without too many more matrix–vector products with $\mathbf{M}_i$. Iteration counts are much smaller for the inexact constraint preconditioner than the block diagonal preconditioner. Overall, using our new preconditioners $\mathbf{L}_M$ and correlated choices of $\widehat{\mathbf{R}}$ result in fewer iterations and matrix–vector products compared to those obtained when using $\mathbf{L}_0$ or $\mathbf{R}_{diag}$ for the Lorenz 96 problem.

7.2. Heat equation with Dirichlet boundary conditions. The second problem of interest that we consider here is the one-dimensional heat equation on the unit line

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}, \tag{7.1}$$

with homogeneous Dirichlet boundary conditions. We discretise (7.1) using the forward Euler method in time and second-order centred differences in space. This means we can write the model evolution in matrix form for a single time model step as $\mathbf{u}^{t+\Delta t} = \mathbf{M}_{\Delta t}\mathbf{u}^t$, where $\mathbf{M}_{\Delta t}$ denotes the application of a single model time-step of length Δt to the heat equation with Dirichlet boundary conditions; this is given by

$$\mathbf{M}_{\Delta t} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1-2r & r & 0 & \cdots & 0 \\ 0 & r & 1-2r & \ddots & & \vdots \\ 0 & 0 & \ddots & \ddots & r & 0 \\ \vdots & \vdots & & r & 1-2r & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix},$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

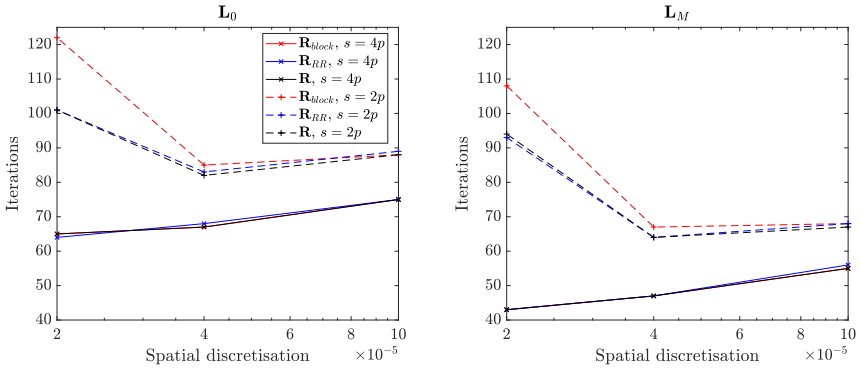

FIG. 7.2. *Experiment B: Number of iterations required for convergence using $\mathbf{L}_0$ (left) and $\mathbf{L}_M$ (right) for different choices of $\widehat{\mathbf{R}}$ within the inexact constraint preconditioner $\boldsymbol{\mathcal{P}}_I$ for the heat equation, for large choices of s. Two observation ratios are shown: $p = \frac{s}{2}$ (dashed line) and $p = \frac{s}{4}$ (solid line). The dimension of the problem ranges from $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{150,000\times 150,000}$ (right) to $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{750,000\times 750,000}$ (left) for $p = \frac{s}{2}$, and $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{135,000\times 135,000}$ (right) to $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{675,000\times 675,000}$ (left) for $p = \frac{s}{4}$. Legend entries correspond to different choices of $\widehat{\mathbf{R}}$ as described in Section 5.*

where $r = \frac{\alpha\Delta t}{(\Delta x)^2}$. The case of non-homogeneous boundary conditions would follow similarly, by applying a source term to the model evolution equation. For the numerical experiments presented here we fix $\alpha = 1$ and vary spatial and temporal resolutions together, setting the ratio $r = \frac{\Delta t}{(\Delta x)^2} = 0.4$ for all experiments.

One advantage of the heat equation test problem is the ability to consider how our new preconditioners scale with problem size. We now consider the best-performing preconditioners for a high-dimensional example, namely the inexact constraint preconditioner for $\mathbf{R}_{block}$, $\mathbf{R}_{RR}$, and $\mathbf{R}$. Experiment B studies the behaviour of the inexact constraint preconditioner for high-dimensional problems. In particular, for $\Delta x = 2 \times 10^{-5}$ the dimension of the full saddle point problem is $750,000 \times 750,000$ for $s = 2p$. In Figure 7.2 we only consider the inexact constraint preconditioner and the three best choices of $\widehat{\mathbf{R}}$. We find that even for a high-dimensional problem the number of iterations is small, with only a modest difference between the results for $\mathbf{R}_{block}$ and $\mathbf{R}$ (as well as $\mathbf{R}_{RR}$). Similarly to the lower dimensional case, using $\mathbf{L}_M$ requires fewer iterations than using $\mathbf{L}_0$. Figure 7.2 also considers two observation networks. We see that using a larger number of observations requires a larger number of iterations to reach convergence, which coincides with the findings of [8] for the unpreconditioned case. However, the qualitative behaviour across the different choices of $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ is the same for both observation networks.

8. Conclusions.

We proposed new preconditioners for the saddle point formulation of the weak-constraint 4D-Var data assimilation problem in the presence of correlated observation errors. Our approach for approximating the model term, $\widehat{\mathbf{L}}$, incorporated model information for the first time. We also proposed a range of approaches that permit inclusion of computationally efficient correlation information within the observation error covariance term, $\widehat{\mathbf{R}}$. In summary:

- We developed new bounds for the eigenvalues of the preconditioned saddle point system in the case of a block diagonal preconditioner.
- We investigated how the constituent terms within the bounds behave for existing and proposed choices of $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$. Including model information via $\mathbf{L}_M$ yields many repeated unit eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^{\top}\mathbf{L}\mathbf{L}_M^{-1}$. Our new approaches yield eigenvalues of this matrix that are frequently bounded above by moderate numbers.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

- We considered two numerical examples: the Lorenz 96 problem and the heat equation. Including model information via $\mathbf{L}_M$ reduced iterations for both problems.

The use of preconditioners that account for correlated observation information led to a significant reduction in iterations for all experiments. For many problems where $\mathbf{R}$ is very ill-conditioned, we would expect the improvements in performance to be even greater than in the experiments presented here. We find that including additional model information in $\mathbf{L}_M$ leads to reduced iterations, but increases the computational expense of each iteration. We therefore suggest that selecting $k = 3$ or $k = 4$ represents a sensible trade-off. Future work for this problem includes developing efficient approximations of $\mathbf{D}$, multi-core implementations of our new preconditioners, and experiments within a full-scale operational NWP system.

Acknowledgements. We thank Adam El-Said for his code for the Lorenz 96 weak constraint 4D-Var assimilation problem. We gratefully acknowledge funding from the Engineering and Physical Sciences Research Council (EPSRC) grant EP/S027785/1 and the London Mathematical Society Research in Pairs scheme.

Appendix: Block preconditioner for $\widehat{\mathbf{R}}$. We state below the algorithm used to apply the block preconditioner $\mathbf{R}_{block}$ for $\mathbf{R}$, as described in Section 5.1.

Algorithm 3: Block preconditioner for $\widehat{\mathbf{R}}$

Inputs: $\mathbf{R}_i$, `pvec` = vector of block sizes, `tol` = tolerance for retaining blocks, `maxsize` = maximum size permitted on a single processor, `numproc` = number of available processors.
Compute `p = sum(pvec)`, `pn = |pvec|`.
Define `pst` = starting index for each new block.
Initialise $\mathbf{R}_{block} = \mathbf{R}_i$.
for $j =$ `1:pn-1`
  Compute scaled Frobenius norm of super-diagonal blocks via
  `normvec(j) = 1/sqrt(pvec(j)*pvec(j+1))*`
  `  norm(R(pst(j):pst(j+1)-1,pst(j+1):pst(j+2)-1),'fro').`
end
Retain blocks where `normvec(j) >= tol`:
for $j =$ `1:pn-1`
  if `normvec(j) < tol`
    Set $\mathbf{R}_{block}$`(pst(j+1):p,1:pst(j+1)-1) = 0.`
    Set $\mathbf{R}_{block}$`(1:pst(j+1)-1,pst(j+1):p) = 0.`
  end
end
if size of largest block $>$ `maxsize`
  Split largest block into two components.
end
if number of distinct blocks $>$ `numproc`
  Combine two smallest adjacent blocks in $\mathbf{R}_{block}$.
elseif number of distinct blocks $<$ `numproc` $- 2$
  Split largest block of $\mathbf{R}_{block}$ into two components.
end

References.

[1] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[2] L. Bergamaschi, J. Gondzio, M. Venturin, and G. Zilli. Inexact constraint preconditioners for linear systems arising in interior point methods. *Computational Optimization and Applications*, 36(2–3):137–147, 2007.
[3] L. Bergamaschi, J. Gondzio, M. Venturin, and G. Zilli. Erratum to: Inexact constraint preconditioners for linear systems arising in interior point methods. *Computational Optimization and Applications*, 49(2):401–406, 2011.
[4] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, Princeton, N.J.; Oxford, UK, 2nd edition, 2009.
[5] M. Bonavita, Y. Trémolet, E. Holm, S. T. K. Lang, M. Chrust, M. Janisková, P. Lopez, P. Laloyaux, P. de Rosnay, M. Fisher, M. Hamrud, and S. English. *A strategy for data assimilation*. European Centre for Medium Range Weather Forecasts, Reading, UK, 2017.
[6] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535, 2018.
[7] E. S. Cooper, S. L. Dance, J. Garcia-Pintado, N. K. Nichols, and P. J. Smith. Observation impact, domain length and parameter estimation in data assimilation for flood forecasting. *Environmental Modelling & Software*, 104:199–214, 2018.
[8] I. Daužickaitė, A. S. Lawless, J. A. Scott, and P. J. Van Leeuwen. Spectral estimates for saddle point matrices arising in weak constraint four-dimensional variational data assimilation. *Numerical Linear Algebra with Applications*, 27(5):e2313, 2020.
[9] A. El Akkraoui, Y. Trémolet, and R. Todling. Preconditioning of variational data assimilation and the use of a bi-conjugate gradient method. *Quarterly Journal of the Royal Meteorological Society: A*, 139(672):731–741, 2013.
[10] A. El-Said. *Conditioning of the weak-constraint variational data assimilation problem for numerical weather prediction*. PhD thesis, University of Reading, 2015.
[11] H. Elman, D. Silvester, and A. Wathen. *Finite Elements and Fast Iterative Solvers*. Oxford University Press, Oxford, UK, 2nd edition, 2014.
[12] M. Fisher, S. Gratton, S. Gürol, Y. Trémolet, and X. Vasseur. Low rank updates in preconditioning the saddle point systems arising from data assimilation problems. *Optimization Methods and Software*, 33(1):45–69, 2018.
[13] M. Fisher and S. Gürol. Parallelization in the time dimension of four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society: B*, 143(703):1136–1147, 2017.
[14] M. A. Freitag and D. L. H. Green. A low-rank approach to the solution of weak constraint variational data assimilation problems. *Journal of Computational Physics*, 357:263–281, 2018.
[15] S. Gratton, S. Gürol, E. Simon, and P. L. Toint. Guaranteeing the convergence of the saddle formulation for weakly constrained 4D-Var data assimilation. *Quarterly Journal of the Royal Meteorological Society: B*, 144(717):2592–2602, 2018.
[16] S. Gratton, A. S. Lawless, and N. K. Nichols. Approximate Gauss–Newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106–132, 2007.
[17] S. Gratton, P. L. Toint, and J. Tshimanga Ilunga. Range-space variants and inexact matrix-vector products in Krylov solvers for linear systems arising from inverse problems. *SIAM Journal on Matrix Analysis and Applications*, 32(3):969–986, 2011.
[18] D. Green. *Model order reduction for large-scale data assimilation problems*. PhD thesis, University of Bath, 2019.
[19] A. Greenbaum, V. Pták, and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 17(3):465–469, 1996.

[20] C. Greif, T. Rees, and D. B. Szyld. GMRES with multiple preconditioners. *SeMA Journal*, 74(2):213–231, 2017.

[21] S. Gürol, A. T. Weaver, A. M. Moore, A. Piacentini, H. G. Arango, and S. Gratton. B-preconditioned minimization algorithms for variational data assimilation with the dual formulation. *Quarterly Journal of the Royal Meteorological Society: B*, 140(679):539–556, 2014.

[22] S. A. Haben. *Conditioning and preconditioning of the minimisation problem in variational data assimilation*. PhD thesis, Department of Mathematics and Statistics, University of Reading, 2011.

[23] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

[24] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[25] Y. A. Kuznetsov. Efficient iterative solvers for elliptic finite element problems on nonmatching grids. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 10(3):187–211, 1995.

[26] E. N. Lorenz. Predictability: a problem partly solved. In *Seminar on Predictability, 4–8 September 1995*, Reading, 1995. ECMWF.

[27] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing*, 21(6):1969–1972, 2000.

[28] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[29] C. C. Paige, M. A. Saunders, S.-C. Choi, D. Orban, U. E. Villa, D. Maddix, and S. Regev. MINRES: Sparse symmetric equations, *software available at* https://web.stanford.edu/group/sol/software/minres/, 2020.

[30] E. M. Pinnington, E. Casella, S. L. Dance, A. S. Lawless, J. I. L. Morison, N. K. Nichols, M. Wilkinson, and T. L. Quaife. Investigating the role of prior and observation error correlations in improving a model forecast of forest carbon balance using Four-dimensional Variational data assimilation. *Agricultural and Forest Meteorology*, 228–229:299–314, 2016.

[31] E. M. Pinnington, E. Casella, S. L. Dance, A. S. Lawless, J. I. L. Morison, N. K. Nichols, M. Wilkinson, and T. L. Quaife. Understanding the effect of disturbance from selective felling on the carbon dynamics of a managed woodland by combining observations with model predictions. *Journal of Geophysical Research: Biogeosciences*, 122(4):886–902, 2017.

[32] F. Rawlins, S. P. Ballard, K. J. Bovis, A. M. Clayton, D. Li, G. W. Inverarity, A. C. Lorenc, and T. J. Payne. The Met Office global four-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society: B*, 133(623):347–362, 2007.

[33] T. Rees. *Preconditioning iterative methods for PDE constrained optimization*. PhD thesis, University of Oxford, 2010.

[34] T. Rees and A. J. Wathen. Preconditioning iterative methods for the optimal control of the Stokes equation. *SIAM Journal on Scientific Computing*, 33(5):2903–2926, 2011.

[35] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[36] D. Simonin, J. A. Waller, S. P. Ballard, S. L. Dance, and N. K. Nichols. A pragmatic strategy for implementing spatially correlated observation errors in an operational system: An application to Doppler radial winds. *Quarterly Journal of the Royal Meteorological Society: B*, 145(723):2772–2790, 2019.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[37] L. M. Stewart, S. L. Dance, and N. K. Nichols. Correlated observation errors in data assimilation. *International Journal for Numerical Methods in Fluids*, 56(8):1521–1527, 2008.

[38] J. M. Tabeart, S. L. Dance, S. A. Haben, A. S. Lawless, N. K. Nichols, and J. A. Waller. The conditioning of least-squares problems in variational data assimilation. *Numerical Linear Algebra with Applications*, 25(5):e2165, 2018.

[39] J. M. Tabeart, S. L. Dance, A. S. Lawless, N. K. Nichols, and J. A. Waller. Improving the condition number of estimated covariance matrices. *Tellus A: Dynamic Meteorology and Oceanography*, 72(1):1–19, 2020.

[40] J. M. Tabeart, S. L. Dance, A. S. Lawless, N. K. Nichols, and J. A. Waller. New bounds on the condition number of the Hessian of the preconditioned variational data assimilation problem. *Numerical Linear Algebra with Applications*, 29(1):e2405, 2022.

[41] P. Tandeo, P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, and Y. Zhen. A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*, 148(10):3973–3994, 2020.

[42] Y. Trémolet. Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society: B*, 132(621):2483–2504, 2006.

[43] Y. Trémolet. Model-error estimation in 4D-Var. *Quarterly Journal of the Royal Meteorological Society: A*, 133(626):1267–1280, 2007.

[44] J. A. Waller, J. García-Pintado, D. C. Mason, S. L. Dance, and N. K. Nichols. Assessment of observation quality for data assimilation in flood models. *Hydrology and Earth System Sciences*, 22(7):3983–3992, 2018.

[45] P. P. Weston, W. Bell, and J. R. Eyre. Accounting for correlated error in the assimilation of high-resolution sounder data. *Quarterly Journal of the Royal Meteorological Society: B*, 140(685):2420–2429, 2014.

[46] A. M. Yaglom. *Correlation Theory of Stationary and Related Random Functions, Volume I: Basic Results*. Springer-Verlag, New York, 1987.