Received: 31 December 2019 | Revised: 14 May 2020 | Accepted: 17 June 2020

DOI: 10.1002/gamm.202000014

ORIGINAL PAPER

Numerical linear algebra in data assimilation

Melina A. Freitag

Institut für Mathematik, Universität Potsdam, Potsdam, Germany

Correspondence
Melina A. Freitag, Institut für Mathematik, Universität Potsdam, Campus Golm, Haus 9, Karl-Liebknecht-Str. 24-25, D-14476 Potsdam, OT, Germany.
melina.freitag@uni-potsdam.de

Funding information
Open access funding enabled and organized by Projekt DEAL

Abstract

Data assimilation is a method that combines observations (ie, real world data) of a state of a system with model output for that system in order to improve the estimate of the state of the system and thereby the model output. The model is usually represented by a discretized partial differential equation. The data assimilation problem can be formulated as a large scale Bayesian inverse problem. Based on this interpretation we will derive the most important variational and sequential data assimilation approaches, in particular three-dimensional and four-dimensional variational data assimilation (3D-Var and 4D-Var) and the Kalman filter. We will then consider more advanced methods which are extensions of the Kalman filter and variational data assimilation and pay particular attention to their advantages and disadvantages. The data assimilation problem usually results in a very large optimization problem and/or a very large linear system to solve (due to inclusion of time and space dimensions). Therefore, the second part of this article aims to review advances and challenges, in particular from the numerical linear algebra perspective, within the various data assimilation approaches.

KEYWORDS

3D-Var, 4D-Var, Bayesian inverse problems, conjugate gradients, GMRES, Kalman filter, Krylov methods, low-rank methods, model order reduction, optimization, preconditioning, sparse linear systems, variational data assimilation

1 | MOTIVATION

Integrating large data sets into sophisticated computational models is one of the big challenges in mathematical sciences of the 21st century. When the computational model arises from a dynamical system, and time dependent observational data of that system are available, then the process of combining the model and the data to obtain a more informed system is called *data assimilation*.

Data assimilation research has been mainly driven by practitioners, initially in the field of numerical weather prediction and ocean modeling [49,50,56,86,87,120,135,139,155,156,164], but nowadays has many more applications in geosciences [38,79,143,171], ecology [130,140], biology [126,151], chemistry [29,66], mechanical engineering [3,47], medicine [68,115], image processing [22,32], as well as human and social sciences [157,159], see also [8] and references therein, with the potential for further utilization in data science and machine learning. In particular, as it becomes easier to make large numbers of relatively accurate observations of a system (we explain later what we mean by a system), a major challenge is how best to use this information to update and refine the model of that system. Only in recent years a more mathematical approach to the theory of data assimilation has been developed, see for example [82,113,119,150,162,176].

Good introductions to Bayesian approaches for inverse problems from the viewpoint of numerical linear algebra are the book [34] and the reference [35].

A model of a system consists of a number of mathematical equations, often (stochastic) partial differential equations that describe the interactions between several variables through certain processes from physics, biology, chemistry, or other applications. Most of the time those equations simplify the dynamics of the system by excluding processes that are considered less important, happen at different scales, or are simply not easy to model. Moreover, parameters in the system may only be known approximately, and the computational model (via discretization of differential equations) results in another error introduced within the process. Even if we did have perfect knowledge of a system, often initial conditions or boundary conditions are not known to high accuracy.

In addition to the model, measurements of the system variables, perhaps indirect, are available at different locations in time and space.

Both the model and the observations obtained through measurements have errors. The data assimilation problem is therefore an inverse problem that uses incomplete and erroneous data and knowledge about a model which is also imperfect, in order to find the best possible estimate of the state of a system (or the best possible approximation of an unknown parameter, in which case we speak of parameter estimation). Using this state estimate one can then use the model to make predictions about the future states of the system, and, as more observations become available, update the state estimate using cycled data assimilation schemes.

Many algorithms for data assimilation have been developed over the past century, we introduce the main ones in the next section. However many challenges remain, in particular as more and more data become available and larger, more complex and higher dimensional problems need to be solved.

This review article will not address all problems arising within data assimilation, but will focus on challenges in data assimilation for numerical linear algebra.

The article is structured as follows: The most common methods for data assimilation are introduced in Section 2: Based on Bayes' theorem we derive variational and sequential data assimilation techniques. Section 3 focuses on the solution to the optimization problem and the linear system arising within variational data assimilation. Approximations, in particular low-rank approximations, and other variants of the Kalman filter are considered in Section 4. Section 5 reviews several dimension reduction approaches for variational and sequential data assimilation methods, and in Section 6 we briefly consider various other aspects, such as the connection between data assimilation, Bayesian inference and Tikhonov regularization. Finally the last two sections give a short survey on data assimilation software and a conclusion. By no means we consider this article a complete survey on data assimilation techniques, we rather focus on a selected methods and approaches where linear algebra plays an important role.

2 | BASIC METHODS AND ALGORITHMS FOR STATE ESTIMATION

The goal of data assimilation is to incorporate measured observations into a model of a dynamical system in order to produce estimates of the current system state (and future system states) which are as accurate as possible. In that sense data assimilation can be defined as an approximation of a true state of a physical system at a given time, by combining time-distributed observations with the dynamical system model in some optimal way.

One can view the data assimilation problem as a Bayesian inference problem. Let $x \in \mathbb{R}^n$ be a model state that we would like to estimate. In Bayesian statistics, we model x as a realization of a random variable (here a random vector), $X : \Omega \to \mathbb{R}^n$. If $\Theta : \Omega \to \mathbb{R}^p$ is another random variable with mean zero modeling the observational noise, then we can also model the observed variable $Y \in \mathbb{R}^p$ as a random variable, defined by

$$Y = h(X) + \Theta,$$

where $h : \mathbb{R}^n \to \mathbb{R}^p$ is a (in general nonlinear) continuous map which models the transformation of the system space to the observation space. X and Θ are assumed to be independent. We would like to infer information about states x given realizations y of Y, which is a Bayesian inverse problem [2,6,12,34,35,57,162].

If we assume that $\Theta \in \mathbb{R}^p$ is a random variable with probability density π, and y as well as x realizations of the random variable Y and X, respectively, then the probability of y given x is given by $\pi_{Y|X}(y|x) = \pi(y - h(x))$. This is often referred to as the data likelihood. Further let $\pi_X(x)$ be the probability density function of X, which describes our prior believes about the distribution of X. Then, by Bayes' formula, $\pi_{X|Y}(x|y)$, the *posterior conditional probability density* function of x given

the observations y is given by

$$\pi_{X|Y}(x|y) = \frac{\pi_{Y|X}(y|x)\pi_X(x)}{\pi_Y(y)} \propto \pi_{Y|X}(y|x)\pi_X(x), \tag{1}$$

where $\pi_Y(y) = \int_{\mathbb{R}^n} \pi_{Y|X}(y|x)\pi_X(x)dx$ is a normalization constant depending only on y (see, eg, [119,149,162]). In general it is hard to obtain the entire probability density $\pi_X(x|y)$, in particular in higher dimensions. However, if we make some assumptions about the probability density functions of the prior and the likelihood, the problem of finding the posterior density can be simplified, which leads to data assimilation algorithms in the traditional sense, which we discuss in the next sections. We will distinguish between variational and sequential data assimilation methods.

2.1 | Variational data assimilation

If the prior density in the Bayesian inference problem (1) is Gaussian, that is $X \sim \mathcal{N}(x^B, B)$ with mean $x^B \in \mathbb{R}^n$ (often called the background vector) and positive definite background error covariance matrix $B \in \mathbb{R}^{n\times n}$, then we have

$$\pi_X(x) = \frac{1}{\sqrt{(2\pi)^n \det B}} \exp\left(-\frac{1}{2}(x - x^B)^T B^{-1}(x - x^B)\right).$$

Here $x \in \mathbb{R}^n$ is the state vector we would like to estimate. If, in addition, the observation error is also Gaussian, that is $\Theta \sim \mathcal{N}(0, R)$ (or $Y \sim \mathcal{N}(h(X), R)$) with positive definite error covariance matrix $R \in \mathbb{R}^{p\times p}$, then the likelihood function can be written as

$$\pi_{Y|X}(y|x) = \frac{1}{\sqrt{(2\pi)^p \det R}} \exp\left(-\frac{1}{2}(y - h(x))^T R^{-1}(y - h(x))\right).$$

and hence

$$\pi_X(x|y) \propto \exp\left(-\frac{1}{2}\|y - h(x)\|^2_{R^{-1}} - \frac{1}{2}\|x - x^B\|^2_{B^{-1}}\right),$$

where $\|z\|^2_{R^{-1}} := z^T R^{-1} z$ is a weighted norm, the so-called Mahalanobis distance to zero. The maximum a posteriori (MAP) estimator for $x \in \mathbb{R}^n$ is then given by

$$\underset{x\in\mathbb{R}^n}{\operatorname{argmin}} J(x), \quad \text{where} \quad J(x) = \left(\frac{1}{2}\|y - h(x)\|^2_{R^{-1}} + \frac{1}{2}\|x - x^B\|^2_{B^{-1}}\right), \tag{2}$$

which is a weighted nonlinear least squares problem. The minimization problem in (2) is what is known as the three-dimensional variational data assimilation problem (*3D-Var*). If $h : \mathbb{R}^n \to \mathbb{R}^p$ is a linear operator, represented by a matrix $H \in \mathbb{R}^{p\times n}$, the solution to (2) can be computed immediately using the (sometimes called Kalman gain) matrix K given by

$$K = BH^T(HBH^T + R)^{-1},$$

or $K = (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}$, using the Sherman-Morrison-Woodbury formula [89], and the minimum in (2) is then given by

$$x^* = x^B + K(y - H(x^B)).$$

This direct solve of the optimization problem (2) is sometimes referred to as optimal interpolation [56].

We note that the minimization problem in (2) is a generalized form of Tikhonov regularization [12,82,102,111,134,138], the most commonly used form of regularization for inverse problem. In standard Tikhonov regularization problems, the operator $R = I$, where I is the identity matrix, $x^B = 0$, $B = \frac{1}{\lambda} I$, where $\lambda > 0$ is a regularization parameter, and often h is linear. The generalized form (2) is a nonlinear Tikhonov regularization in a weighted inner

product space (see, [35,82,175]). In recent years the dynamic aspect of inverse problems (naturally leading to a variety of data assimilation problems) has become of interest in the inverse problems community, we especially refer to [112,158] and references therein.

In practice the matrices involved in the computation of the Kalman gain are very large and direct inversion is not feasible, often it is not even possible to store the full matrices. In addition, if the operator h is nonlinear, then an iterative optimization procedure is required in order to solve the minimization problem (2), we will discuss more details about this optimization in Section 3.

In variational data assimilation one uses a descent algorithm in the direction of the gradient and an adjoint approach for the computation of the gradient in order to solve the minimization problem.

The problem (2) is a stationary one, there is no time-dependence. Instead of a state x fixed in time, let us consider a set of state estimates $x = [x_0^T, \dots, x_N^T]^T$, where $x_i \in \mathbb{R}^n$ refers to a state at time t_i. Define a new observation operator $\mathcal{H} : \mathbb{R}^{n(N+1)} \to \mathbb{R}^{p(N+1)}$, with $\mathcal{H} : [x_0^T, \dots, x_N^T]^T \to [\mathcal{H}_0(x_0)^T, \dots, \mathcal{H}_N(x_N)^T]^T$ and let $y = [y_0^T, \dots, y_N^T]^T$, where $y_i \in \mathbb{R}^p$ denotes a set of observations at time t_i, $i = 0, \dots, N$. Furthermore, let $R \in \mathbb{R}^{p(N+1)\times p(N+1)}$ now be a block diagonal matrix with $R_i \in \mathbb{R}^{p\times p}$, $i = 0, \dots, N$, on the diagonal blocks. Again assuming the observation errors are distributed according to a normal distribution with error covariance matrix R_i we define the cost function

$$J(x_0) = \frac{1}{2}\|y - \mathcal{H}(x)\|^2_{R^{-1}} + \frac{1}{2}\|x_0 - x_0^B\|^2_{B^{-1}} = \frac{1}{2}\sum_{i=0}^{N}\|y_i - \mathcal{H}_i(x_i)\|^2_{R_i^{-1}} + \frac{1}{2}\|x_0 - x_0^B\|^2_{B^{-1}}. \tag{3}$$

In addition we introduce a discrete model for the evolution of the underlying physical system from time t_i to time t_{i+1}, described by the dynamical system equations

$$x_{i+1} = \mathcal{M}_i(x_i),$$

where $\mathcal{M}_i : \mathbb{R}^n \to \mathbb{R}^n$ can be time-dependent and nonlinear, describing the evolution of the dynamical system. Usually this forward operator requires the solution of a time-dependent partial differential equation. The problem

$$\underset{x_0\in\mathbb{R}^n}{\operatorname{argmin}} J(x_0), \quad \text{where} \quad J(x_0) = \frac{1}{2}\sum_{i=0}^{N}\|y_i - \mathcal{H}_i(x_i)\|^2_{R_i^{-1}} + \frac{1}{2}\|x_0 - x_0^B\|^2_{B^{-1}}, \tag{4}$$

$$\text{subject to the nonlinear model dynamics} \quad x_{i+1} = \mathcal{M}_i(x_i) \tag{5}$$

is a constrained optimization problem and is called *4D-Var*, four dimensional variational data assimilation. Here, the subscripts refer to the time index. Note that in (4) the regularization term only involves the initial state x_0 at time t_0. The minimization provides the initial condition of the model that most closely fits the data.

Figure 1 illustrates the difference between 4D-Var and 3D-Var (for illustration purposes the state dimension is $n = 1$). The 3D-Var problem finds the best estimate using an observation and a background state x^B at one specific time point (here at the initial time), the best estimate is called $x^A_{\text{3D-Var}}$. In 4D-Var observations are obtained throughout a time window—often called the assimilation window—and the optimization is done over a time window. The best estimate, taking into account the background state, that is the previous forecast and the observations within the assimilation window is computed, where we have to take the dynamical forecast model into account. The best estimate at the initial state is then $x^A_{\text{4D-Var}}$, which of course is not necessary the same as $x^A_{\text{3D-Var}}$. Both estimates can be used to evolve the model forward and to obtain a better (updated) forecast.

The cost function in (2) or (4) is minimized using an iterative method (eg, steepest descent, conjugate gradient, a quasi-Newton method, etc. [141]). We choose $x_0^{(0)}$ (often $x_0^{(0)} = x_0^B$) and update

$$x_0^{(\ell+1)} = x_0^{(\ell)} + \alpha^{(\ell)} d^{(\ell)}, \quad \ell = 0, 1, 2, \dots,$$

where $\alpha^{(\ell)}$ is a damping parameter obtained through line search, for example, and ℓ is the iteration index. The search direction is $d^{(\ell)} = -\nabla J(x_0^{(\ell)})$ for gradient descent, or $d^{(\ell)} = -(\nabla^2 \tilde{J}(x_0^{(\ell)}))^{-1}\nabla J(x_0^{(\ell)})$ for a quasi-Newton method (where $\nabla^2\tilde{J}(x_0^{(\ell)})$ is an approximation of the Hessian of $J(x_0^{(\ell)})$, or Jacobian matrix of $\nabla J(x_0^{(\ell)})$).

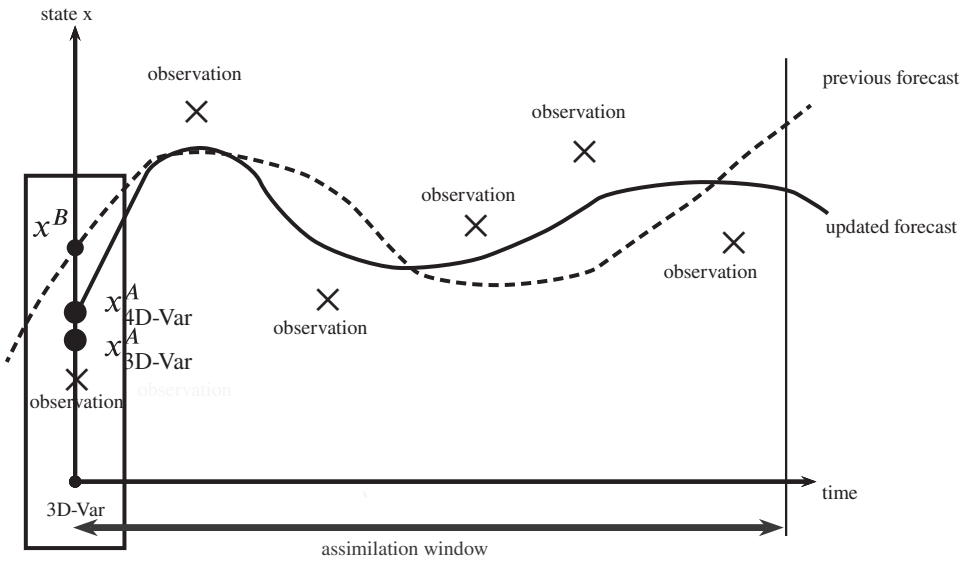

FIGURE 1 Illustration of 3D-Var and 4D-Var for state estimation in one dimension

Clearly, 4D-Var (4) and (5) is a constrained optimization problem, and the gradient $\nabla J(x_0^{(\ell)})$ is obtained via an adjoint approach (see [8]): We introduce Lagrange multipliers $\lambda_i \in \mathbb{R}^n$ at time t_i, $i=1, \dots, N$ and define the Lagrangian

$$\mathcal{L}(x_i, \lambda_i) = J(x_0) + \sum_{i=0}^{N-1} \lambda_{i+1}^T (x_{i+1} - \mathcal{M}_i(x_i)).$$

Necessary conditions for the minimum of the 4D-Var cost function subject to the constraint are then found by taking the variations of $\mathcal{L}$ with respect to λ_i and x_i (KKT conditions, see [141])

$$\begin{aligned}
\frac{\partial \mathcal{L}(x_i, \lambda_i)}{\partial x_0} &= B^{-1}(x_0 - x_0^B) + H_0^T R_0^{-1}(\mathcal{H}_0(x_0) - y_0) - M_0^T \lambda_1 = 0, \\
\frac{\partial \mathcal{L}(x_i, \lambda_i)}{\partial x_i} &= H_i^T R_i^{-1}(\mathcal{H}_i(x_i) - y_i) - M_i^T \lambda_{i+1} + \lambda_i = 0, \quad i = 1, \dots, N, \\
\frac{\partial \mathcal{L}(x_i, \lambda_i)}{\partial \lambda_i} &= x_i - \mathcal{M}_{i-1}(x_{i-1}) = 0 \quad i = 1, \dots, N,
\end{aligned}$$

where $M_i \in \mathbb{R}^{n \times n}$ and $H_i \in \mathbb{R}^{p \times n}$ are the Jacobians of the forward operator $\mathcal{M}_i$ and the observation operator $\mathcal{H}_i$, evaluated at x_i, respectively, that is

$$M_i = \frac{\partial \mathcal{M}_i}{\partial x}(x_i), \quad H_i = \frac{\partial \mathcal{H}_i}{\partial x}(x_i).$$

The adjoint equations for the adjoint variables λ_i, $i=0, \dots, N+1$, that measure the sensitivity of the cost function to changes in x_i, are then given by

$$\lambda_{N+1} = 0 \tag{6}$$

$$\lambda_i = M_i^T \lambda_{i+1} - H_i^T R_i^{-1}(\mathcal{H}_i(x_i) - y_i), \quad i = N, \dots, 0. \tag{7}$$

They provide an efficient method to compute the gradient of the objective function (4), which is then given by

$$\nabla J(x_0) = -\lambda_0 + B^{-1}(x_0 - x_0^B).$$

The general method for solving the 4D-Var optimization problem is sketched in Algorithm 1.

Algorithm 1. 4D-Var

Input: error covariance matrices R_i and B, routines to apply model and observation operators $\mathcal{M}_i$ and $\mathcal{H}_i$ and their linearizations M_i and H_i, respectively, and observations y_i for $i = 0, \dots, N$, maximum number of iterations $\ell_{\max}$.
Initialize the iteration $\ell = 0$ and $x_0^{(0)} = x_0^B$.
while $\|\nabla J(x_0^{(\ell)})\| < \varepsilon$ or $\ell \leq \ell_{\max}$ **do**
  Compute cost function $J(x_0^{(\ell)})$ using the forward model.
  Compute $\nabla J(x_0^{(\ell)})$ using the adjoint equations.
  Apply descent method (gradient descent, Newton, quasi-Newton, …) to compute descent direction $d^{(\ell)}$.
  Update the initial condition $x_0^{(k+1)} = x_0^{(\ell)} + \alpha^{(\ell)} d^{(\ell)}$.
  Set $\ell = \ell + 1$.
end while

The problem in (4) and (5) is often called *strong constraint 4D-Var*. In the presence of uncertainty we can model the imperfect state dynamics by

$$x_{i+1} = \mathcal{M}_i(x_i) + \eta_i,$$

where $\eta_i \sim \mathcal{N}(0, Q_i)$ is the model error which is assumed to be Gaussian with error covariance Q_i, uncorrelated in time and uncorrelated with the background and observation errors. The relaxation of the strong constraint is commonly used in sequential data assimilation as we will see in Section 2.2. For variational data assimilation a weak constraint was proposed in [156], however, due to high computational costs is not used heavily in practice. In the past decades, with the availability of increasing computing power, there has been greater interest in this method [77,80,90,168]. The cost function for weak constraint 4D-Var is given by

$$J(x) = \frac{1}{2}\sum_{i=1}^{N} \|y_i - \mathcal{H}_i(x_i)\|^2_{R_i^{-1}} + \frac{1}{2}\sum_{i=0}^{N-1} \|x_{i+1} - \mathcal{M}_i(x_i)\|^2_{Q_{i+1}^{-1}} + \frac{1}{2}\|x_0 - x_0^B\|^2_{B^{-1}}, \tag{8}$$

where $x = [x_0^T, \dots, x_N^T]^T$, and x_i is the model state at time step t_i, $i = 0, \dots, N$. The resulting optimization problem

$$\underset{x \in \mathbb{R}^{n(N+1)}}{\operatorname{argmin}} J(x) \tag{9}$$

can become extremely large (in three spatial and the additional time dimension) and computationally expensive to solve, as $\mathcal{M}_i$ usually arises from a the discretization of a nonlinear partial differential equation. We will discuss more detailed solution approaches to (9) in Section 3.

Note that the variational approach assumes that prior and likelihood have Gaussian distributions. As such the posterior is also Gaussian if the observation operators h (for 3D-Var) or $\mathcal{H}$ (for 4D-Var, in which case it also includes the model operator) are linear. These are quite strong assumptions and, by minimizing the cost function J in (2), (3), or (4) and (5) we only compute the posterior mean (the MAP estimator).

We are able to approximate the posterior covariance matrix A by computing the inverse of the Hessian, $A = (\nabla^2 J(x))^{-1}$. This covariance matrix is exact for linear models and Gaussian distributions, in which case it is given by $A = (B^{-1} + \mathrm{H}^T \mathrm{R}^{-1} \mathrm{H})^{-1}$, with $\mathrm{H} = H$ and $\mathrm{R} = R$ for 3D-Var, and $\mathrm{H} = [H_0, H_1 M_0, H_2 M_1 M_0, \dots, H_N M_{N-1} \dots M_0]^T$ and $\mathrm{R} = \mathrm{diag}(R_0, R_1, \dots, R_N)$ for 4D-Var, respectively (see more details about this at the end of Section 2.2 and in reference [84]). If the models are nonlinear or the distributions non-Gaussian, the inverse of the Hessian $(\nabla^2 J(x))^{-1}$ is only an approximation of the posterior covariance matrix.

To conclude this section we emphasize that variational data assimilation is essentially based on optimal control theory, that is we minimize a cost function subject to a constraint arising from the state dynamics. The minimization requires techniques from numerical optimization. For more insight into the relation of variational data assimilation and PDE-constrained optimization we refer to [8,76,98].

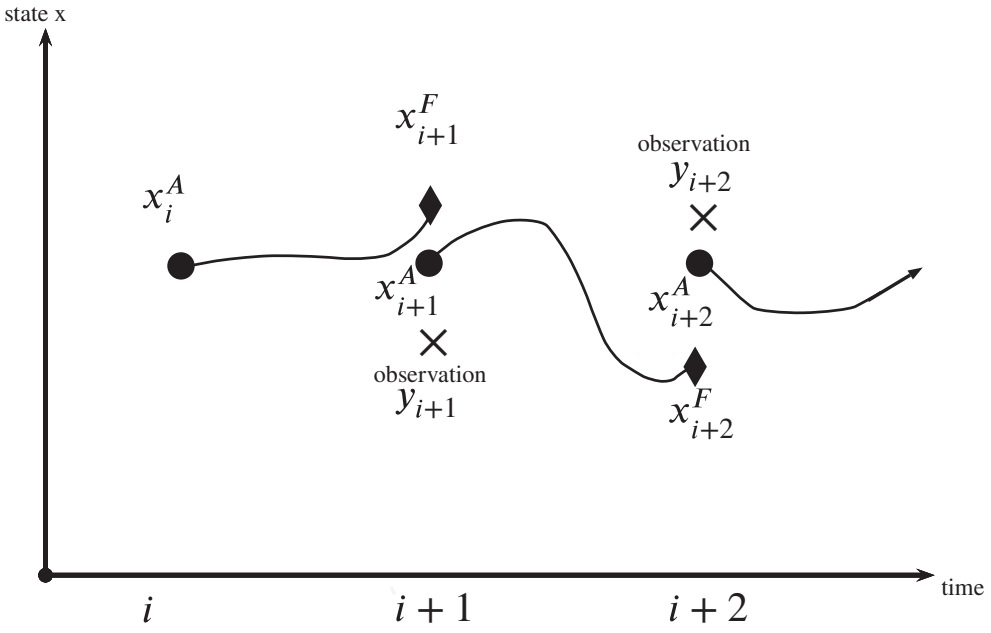

FIGURE 2 Illustration of sequential data assimilation for the Kalman filter in one dimension.

2.2 | Sequential data assimilation and Kalman filters

In sequential data assimilation we correct the model state estimate whenever observations are available, that is we incorporate the observations sequentially. Recall the Bayesian inference problem (1) with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}^p$ observations at time t_i, for $i = 0, \dots, N$. Given a prior uncertainty $\pi_X(x_i)$, find the updated conditional posterior uncertainty $\pi_{X|Y}(x_i|y_i)$ of x_i, given the observation y_i. Using the likelihood $\pi_{Y|X}(y_i|x_i)$ and Bayes' formula we would like to compute

$$\pi_{X|Y}(x_i|y_i) \propto \pi_{Y|X}(y_i|x_i)\pi_X(x_i). \tag{10}$$

Again, ideally we wish to compute the the entire probability density function $\pi_{X|Y}(x_i|y_i)$ at every time step t_i. We consider a discrete-time linear system dynamics,

$$x_{i+1} = M_i x_i + w_i, \tag{11}$$

$$y_i = H_i x_i + v_i, \tag{12}$$

where $w_i \sim \mathcal{N}(0, Q_i)$ and $v_i \sim \mathcal{N}(0, R_i)$ represent model and observation/measurement errors and are assumed to be independent, and Gaussian with error covariance matrices $Q_i \in \mathbb{R}^{n\times n}$ and $R_i \in \mathbb{R}^{p\times p}$ respectively. Hence the errors are uncorrelated in time and between each other. Moreover, model and observation operators are assumed to be linear here, however, extensions to nonlinear models exist.

The Kalman filter (as well as many other sequential data assimilation schemes) then follow two steps: A forecast or prediction step and an analysis or correction step, illustrated in Figure 2. At time step t_i we have an analysis x_i^A available, which we assume to be the best estimate of the state at time t_i (we have incorporated all our knowledge about observations and previous forecasts at time t_i). Then the state dynamics are evolved forward in time using (11) producing a forecast x_{i+1}^F, which becomes the background state at step $i+1$. We also have a set of observations y_{i+1} available which we combine with x_{i+1}^F in order to get the best estimate of the state, a so-called analysis, x_{i+1}^A. This analysis is obtained using the best linear unbiased estimate (BLUE), see [107]. This is again used to evolve the state forward using the model dynamics.

Let $x_i^t \in \mathbb{R}^n$ be the unknown exact true state of the system. Then we define the error in the forecast at time t_i by $e_i^F = x_i^F - x_i^t$, and similar, the error in the analysis at time t_i by $e_i^A = x_i^A - x_i^t$. Moreover, the forecast and analysis error

covariance matrices can then be expressed as

$$P_i^F = \mathrm{cov}(e_i^F) = \mathbb{E}[e_i^F (e_i^F)^T] \quad \text{and} \quad P_i^A = \mathrm{cov}(e_i^A) = \mathbb{E}[e_i^A (e_i^A)^T],$$

respectively.

2.2.1 | Forecast/predictor step

Given a previous analysis state estimate x_i^A at time t_i, an estimate of x_{i+1}^t at time t_{i+1} is given by the application of the model dynamics (11)

$$x_{i+1}^F = M_i x_i^A.$$

For the corresponding error covariance matrix $P_{i+1}^F = \mathbb{E}[e_{i+1}^F (e_{i+1}^F)^T]$ we observe $e_{i+1}^F = M_i e_i^A + w_i$. Since the model is linear and the analysis error e_i^A and model error w_i are assumed to be uncorrelated we obtain (using properties of the expected value and $\mathbb{E}[w_i w_i^T] = Q_i$)

$$P_{i+1}^F = M_i P_i^A M_i^T + Q_i. \tag{13}$$

2.2.2 | Analysis/corrector step

Given an a-priori estimate x_i^F and observations y_i at time step i, the goal of the analysis step is to compute the optimal a-posteriori estimate, x_i^A as a linear combination of x_i^F and y_i of the form

$$x_i^A = x_i^F + K_i(y_i - H_i x_i^F), \tag{14}$$

where $K_i \in \mathbb{R}^{n\times p}$ is called the Kalman gain matrix [114] and the general form of (14) arises from the assumption that the estimate x_i^A of the true vector x_i^t should be both linear and unbiased. The a-posteriori error covariance is given by $P_i^A = \mathbb{E}[e_i^A (e_i^A)^T] = \mathbb{E}[(K_i(-H_i e_i^F + v_i) + e_i^F)(K_i(-H_i e_i^F + v_i) + e_i^F)^T]$, where we have used the observation equation (12). With the definition of $P_i^F := \mathbb{E}[e_i^F (e_i^F)^T]$ and $\mathbb{E}[v_i v_i^T] =: R_i$, we obtain

$$P_i^A = (I - K_i H_i) P_i^F (I - K_i H_i)^T + K_i R_i K_i^T, \tag{15}$$

where we have also used that the forecast error e_i^F and observation error v_i are uncorrelated, random variables, that is $\mathbb{E}[e_i^F v_i^T] = 0$.

The Kalman gain matrix K_i in (15) is chosen to minimize the a-posterior variance $\mathrm{tr}(P_i^A)$. We evaluate $\frac{\partial P_i^A}{\partial K_i} = 0$, and using results from matrix differential calculus we obtain the Kalman gain

$$K_i = P_i^F H_i^T (H_i P_i^F H_i^T + R_i)^{-1}, \tag{16}$$

for details of the derivation, see [8,13,107]. Substituting K_i into (15) then yields

$$P_i^A = (I - K_i H_i) P_i^F. \tag{17}$$

Note that, for P_i^F small (or zero) the Kalman gain is also small (or zero) and the a-posteriori estimate x_i^A is heavily geared towards x_i^F. Similarly, if the observation error covariance matrix R_i is zero (for perfect observations) the estimate x_i^A is steered towards the observations (in particular, for H_i square and nonsingular $x_i^A = H_i^{-1} y_i$). We remark that the error covariance matrices in the Kalman filter equations satisfy discrete algebraic Riccati equations [100,117].

Clearly the Kalman filter has some shortcomings, as it is only optimal for linear models and observation operators, and Gaussian error statistics (in which case the mean and error covariances, which are propagated in the Kalman filter, are sufficient to describe the probability density function of the state estimates).

Several methods have been proposed to overcome these issues, some we mention here, for more details we refer to [8]. The *extended Kalman filter* (EKF) applies to systems of the form

$$x_{i+1} = \mathcal{M}_i(x_i) + w_i,$$
$$y_i = \mathcal{H}_i(x_i) + v_i,$$

where $\mathcal{M}_i$ and $\mathcal{H}_i$ are nonlinear. The nonlinearities are dealt with in the filter equations by linearizing the model and observation operator about x_i^F and x_i^A, in the computations of the Kalman gain and error covariance matrices, that is $M_i = \frac{\partial \mathcal{M}_i}{\partial x}(x_i^A)$, $H_i = \frac{\partial \mathcal{H}_i}{\partial x}(x_i^F)$. For computing the analysis state estimate and the forecast state estimate the nonlinear operators are used. We obtain an approximation to the best linear unbiased estimate, see [85,110]. An algorithmic description of the EKF is given in Algorithm 2.

Algorithm 2. Extended Kalman filter

Input: error covariance matrices R_i and Q_i, routines to apply model and observation operators $\mathcal{M}_i$ and $\mathcal{H}_i$ and their linearizations M_i and H_i, respectively, and observations y_i for $i = 0, \dots, N$.
Initialize the system state x_0^F and the corresponding error covariance matrix P_0^F.
for $i = 0, \dots, N$ **do**
  Compute Kalman gain $K_i = P_i^F H_i^T (H_i P_i^F H_i^T + R_i)^{-1}$.
  Compute state estimate $x_i^A = x_i^F + K_i(y_i - \mathcal{H}_i(x_i^F))$.
  Compute error covariance estimate $P_i^A = (I - K_i H_i) P_i^F$.
  Compute the forecast state $x_{i+1}^F = \mathcal{M}_i(x_i^A)$.
  Compute the forecast error covariance $P_{i+1}^F = M_i P_i^A M_i^T + Q_i$.
end for

There are a lot of similarities between the Kalman filter discussed in this section and variational data assimilation discussed in Section 2.1. In particular, one can interpret the analysis/corrector step in the Kalman filter in a variational form: the Kalman filter state estimate x_i^A from (14) with the Kalman gain (16) exactly represent the equations we obtained for computing the solution to the 3D-Var problem in Section 2.1. Hence, for the variational formulation of the Kalman filter analysis step we have

$$x_i^A = \text{argmin}_{x_i \in \mathbb{R}^n} J_i(x_i), \quad \text{where} \quad J_i(x_i) = \left(\frac{1}{2} \|y_i - H_i(x_i)\|^2_{R_i^{-1}} + \frac{1}{2} \|x_i - x_i^F\|^2_{(P_i^F)^{-1}} \right), \tag{18}$$

which is a quadratic problem for a linear observation operator H_i and hence has a unique solution which can be obtained by setting the gradient $\nabla J_i(x_i) = 0$. For a nonlinear observation operator the solution may not be unique. The Hessian of this cost function (for linear observation operator H_i) is given by

$$\nabla^2 J_i(x_i) = H_i^T R_i^{-1} H_i + (P_i^F)^{-1},$$

and, with the Sherman-Morrison-Woodbury formula we have

$$\nabla^2 J_i(x_i)^{-1} = P_i^F - P_i^F H_i^T (H_i P_i^F H_i^T + R_i)^{-1} H_i P_i^F,$$

which is precisely the a-posteriori error covariance matrix P_i^A given in (17). Therefore, the inverse of the Hessian is equal to the posterior covariance (in the linear case with Gaussian errors). For nonlinear problems, $\nabla^2 J_i(x_i)^{-1}$ is an approximation to the posterior covariance. Solving (18) iteratively is more advantageous than computing the Kalman gain directly for very large problems with sparsity structure. More details on this idea can be found in [13].

To complete this section we add a remark about Kalman smoothers. Filtering algorithms make use of observations as they become available and provide the best estimate of a state given all past information and the current observation. Smoothing algorithms provide the best estimate of a system, using past, present and future information. So the Kalman smoother is equivalent to the Kalman filter at the final time step. Moreover, it can be shown that for linear problems with linear observation operator and model dynamics, and the same initial background error covariance, the Kalman filter and 4D-Var result in the same state estimate for the same time step (when the same observations have been used), see [75,82,127].

In the next sections we give an overview of important approaches and contributions from numerical linear algebra to solve variational and sequential data assimilation problems efficiently and provide an outlook for challenges ahead.

3 | SOLUTIONS TO THE OPTIMIZATION PROBLEM ARISING IN VARIATIONAL DATA ASSIMILATION

Variational data assimilation leads to large PDE-constrained optimization problems. We will first concentrate on 4D-Var (4) and (5) and later discuss weak constraint 4D-Var (8).

3.1 | Incremental 4D-Var and Gauss-Newton method

Since the full nonlinear minimization (4) and (5) is difficult to solve, special algorithms for optimization and linear algebra are required. In geoscience applications an incremental approach [50] was proposed, which is merely a Gauss-Newton method for nonlinear least squares problems [94,121,122]. In incremental variational data assimilation the solution to the nonlinear optimization problem is approximated by a sequence of minimizations of quadratic (and hence convex) cost functions, which are obtained by linearizing both the model and the observation operators. Let $x_0^{(\ell)}$ be the ℓth estimate to the solution at x_0. We linearize the cost function (4) around the model trajectory from this estimate: we linearize the model and observation operators about $x_0^{(\ell)}$ and then obtain the next iterate by the increment $\delta x_0^{(\ell)}$,

$$x_0^{(\ell+1)} = x_0^{(\ell)} + \delta x_0^{(\ell)}. \tag{19}$$

We use this expansion and substitute it into the nonlinear cost function (4), which we then linearize about the model trajectory obtained from $x_0^{(\ell)}$. We then see that the increment $\delta x_0^{(\ell)}$ is obtained by minimizing the quadratic problem (the incremental cost function)

$$\tilde{J}^{(\ell)}(\delta x_0^{(\ell)}) = \frac{1}{2}\sum_{i=0}^{N} \|H_i \delta x_i^{(\ell)} - d_i^{(\ell)}\|^2_{R_i^{-1}} + \frac{1}{2}\|\delta x_0^{(\ell)} - (x_0^B - x_0^{(\ell)})\|^2_{B^{-1}}, \tag{20}$$

where $d_i^{(\ell)} = y_i - \mathcal{H}_i(x_i^{(\ell)})$ and $x_i^{(\ell)}$ is the nonlinear trajectory computed from the current estimate at the initial time $x_0^{(\ell)}$, using the nonlinear model trajectory. H_i is the linearization of the observation operator $\mathcal{H}_i$ about $x_i^{(\ell)}$. The perturbations $\delta x_i^{(\ell)}$ satisfy the linearized constraint

$$\delta x_{i+1}^{(\ell)} = M_i(\delta x_i^{(\ell)}),$$

where M_i is the linear solution operator of the nonlinear model $\mathcal{M}_i$ linearized around the nonlinear trajectory. We therefore obtain a so-called inner-outer iterative method. The outer iteration (with iteration index ℓ) is represented by the update of the nonlinear trajectory (19). The minimization of the quadratic cost function (20) is the inner iteration, it essentially amounts to the solution of a large linear system. At each iteration the forward model, and its linearization, are evaluated in order to compute the cost function (20), and the adjoint model (7) is applied in order to compute the gradient of the cost function.

Incremental variational data assimilation can be shown to be a version of the Gauss-Newton method applied to the original nonlinear cost function (3). In order to illustrate this, consider a general nonlinear least squares

problem

$$\min_x \Phi(x) = \frac{1}{2} f(x)^T f(x) = \frac{1}{2} \|f(x)\|^2, \tag{21}$$

with $f : \mathbb{R}^n \to \mathbb{R}^q$ a twice continuously differentiable function $[f_1(x), \ldots, f_q(x)]$ and $\|\cdot\|$ denoting the Euclidean norm. Let $G(x) = f'(x)$ be the $q \times n$ Jacobian of $f(x)$. The gradient and Hessian of $\Phi(x)$ are then given by

$$\nabla\Phi(x) = G(x)^T f(x), \quad \nabla^2\Phi(x) = G(x)^T G(x) + \sum_{j=1}^{q} f_j(x) \nabla^2 f_j(x).$$

The Gauss-Newton method discards the expensive second term in the Hessian when applying Newton's method to $\nabla\Phi(x) = G(x)^T f(x) = 0$.

Algorithm 3. Gauss-Newton method

Input: Routines to compute $f(x)$ and its Jacobian $G(x)$, maximum number of iterations $\ell_{\max}$.
Initialize the iteration $\ell = 0$ and $x^{(0)} = x_0$
while $\|G(x^{(\ell)})^T f(x^{(\ell)})\| < \varepsilon$ or $\ell \leq \ell_{\max}$ **do**
  Solve $G(x^{(\ell)})^T G(x^{(\ell)}) \delta x^{(\ell)} = -G(x^{(\ell)})^T f(x^{(\ell)})$.
  Update $x^{(\ell+1)} = x^{(\ell)} + \delta x^{(\ell)}$.
  Set $\ell = \ell + 1$.
end while

The Gauss-Newton method is described in Algorithm 3. Note that the second step of the Algorithm is equivalent to solving the linearized least squares problem

$$\min_{s \in \mathbb{R}^n} \|G(x^{(\ell)})s + f(x^{(\ell)})\| \tag{22}$$

at each iteration ℓ. If we define

$$f(x_0) = \begin{bmatrix} B^{-1/2}(x_0 - x_0^B) \\ R_0^{-1/2}(y_0 - \mathcal{H}_0(x_0)) \\ \vdots \\ R_N^{-1/2}(y_N - \mathcal{H}_N(x_N)) \end{bmatrix}$$

subject to $x_{i+1} = \mathcal{M}_i(x_i)$, then the general cost function (21), with $x = x_0$ is equivalent to the 4D-Var cost function (4)-(5), where here $q = n + (N+1)p$. When applying the Gauss-Newton method to (21), then the linearized least squares problem (22) is equivalent to finding the solution to the quadratic cost function (20) with $x = x_0$. More details and also suitable stopping criteria for the inner-outer iteration arising within the Gauss-Newton method were discussed in [94,121-123].

Besides Gauss-Newton, there are of course standard methods to solve nonlinear least squares problems, such as Levenberg-Marquardt and Quasi-Newton methods (in particular BFGS), see, for example [76,141].

3.2 | The inner iteration and preconditioning

The quadratic cost function (20) (see also (22)) within the Gauss-Newton (incremental) approach can be minimized using a conjugate gradient (CG) method [88,89,104,141]. As the problem is usually very large, in practice, the inner loop, the CG iteration, is applied to a system with lower spatial resolution and with a preconditioner [76]. One of the earliest

approaches is the multilevel setting in [50], where the quadratic cost function is replaced by a lower resolution model, in order to obtain a multiresolution scheme.

In addition, two-level preconditioning is usually applied within 4D-Var in the following way. We consider the cost function

$$J(x_0) = \frac{1}{2}\|y - \mathcal{H}(x)\|^2_{R^{-1}} + \frac{1}{2}\|x_0 - x_0^B\|^2_{B^{-1}}, \tag{23}$$

where $\mathcal{H} : [x_0^T, \dots, x_N^T]^T \to [\mathcal{H}_0(x_0)^T, \dots, \mathcal{H}_N(x_N)^T]^T$ for 4D-Var and $\mathcal{H} : x_0 \to \mathcal{H}_0(x_0)$, where $\mathcal{H}_0(x_0) = h(x_0)$ for 3D-Var, respectively. The formulation in (23) aims to treat 3D-Var and 4D-Var simultaneously.

At each step of the Gauss-Newton process an equation of the form $G(x^{(\ell)})^T G(x^{(\ell)})\delta x^{(\ell)} = -G(x^{(\ell)})^T f(x^{(\ell)})$ needs to be solved, which results in

$$(B^{-1} + H^T R^{-1} H)\delta x^{(\ell)} = B^{-1}(x_0^B - x^{(\ell)}) + H^T R^{-1}(y - \mathcal{H}(x^{(\ell)})),$$

where H depends on $x^{(\ell)}$. Note that H is the linearization of $\mathcal{H}$ at $x^{(\ell)}$. For 4D-Var H includes the model operator, $H = [H_0, H_1 M_0, \dots, H_N M_{N-1} \dots M_0]^T \in \mathbb{R}^{(N+1)p\times n}$ and $R = \mathrm{diag}(R_0, R_1, \dots, R_N) \in \mathbb{R}^{(N+1)p\times(N+1)p}$, for 3D-Var $H = H_0$ and $R = R_0$. We exclude the superscript ℓ for ease of notation. The Hessian of the 4D-Var cost function at the ℓth linearized problem is given by $B^{-1} + H^T R^{-1} H$ [83].

The first level preconditioning employs a linear change in variables,

$$\delta x = L\delta\tilde{x}, \quad \text{where} \quad B = LL^T,$$

that is, the Cholesky factor of the background error covariance matrix. The Hessian of the cost function then becomes $I + L^T H^T R^{-1} H L$, which ensures that the smallest eigenvalue of the transformed Hessian is one. Moreover, since the rank of $L^T H^T R^{-1} H L$ is smaller than the dimension of the system, there are many unit eigenvalues. Note that another option is to employ a transformation with B to obtained the transformed Hessian $I + BH^T R^{-1} H$, which is no longer symmetric (but the use of this transformation may be necessary if the factorization of B is not available). Both transformed Hessians share the same eigenvalues [65]; they are all greater than or equal to one. Detailed analysis on the conditioning of the Hessian can be found in [101,163].

At a second preconditioning level the spectrum of the (symmetric) Hessian is used. This utilizes the fact that a few of the dominant eigenvalues and corresponding eigenvectors can be obtained using the Lanczos method [54,118], and, one can compute those Hessian eigenpairs within the CG iteration itself [89]. After k steps of the CG algorithm, a few approximate leading eigenpairs (λ_i, w_i), $i = 1, \dots, k$ of the Hessian are available and one can approximate the Hessian by

$$C = I + \sum_{i=1}^{k} (\lambda_i - 1) w_i w_i^T, \tag{24}$$

where k is much smaller than the dimension of the linear system. Hence, after one step of the Gauss-Newton method (the outer iteration), approximations to the Hessian, C, are available for the next step of the Gauss-Newton iteration, and preconditioners C^{-1}, or $C^{-\frac{1}{2}}$ for H can be obtained by replacing λ_i in (24) by $1/\lambda_i$ or $1/\sqrt{\lambda_i}$, respectively, see [4,133,167]. In [65] this procedure was extended to nonsymmetric Hessians by using the bi-conjugate gradient method.

The work in [28] takes up the idea of approximating the Hessian by eigenpairs obtained from the Lanczos procedure. Matrix-vector products with the Hessian are expensive (since they require the evaluation of the linearized forward and adjoint models), so even obtaining the limited memory representation (24) may be computationally infeasible for large k. Therefore [28] propose a multilevel version of the limited memory Hessian, where the eigenvalue decompositions are obtained from several coarser levels and fed through to the finer level, which enhances the algorithm in terms of reducing the number of matrix vector products at the fine grid levels. In addition, multigrid solvers and multigrid preconditioners for the solution of the variational data assimilation problem were considered recently in [59,95].

The limited memory preconditioners (LMP) discussed above are often referred to as spectral LMP, as they use the spectral information of the Hessian approximation (24). More general versions of LMP were investigated in [93,170]. In

the latter the authors also show the equivalence of a certain reduced version of 4D-Var and the SEEK filter, discussed in Section 4, and use this equivalence to accelerate the convergence of the Gauss-Newton method.

3.3 | The dual formulation of 4D-Var

As observed in the previous paragraph, both for 3D-Var and for 4D-Var, each inner iteration of the Gauss-Newton method essentially attempts to minimize the cost function

$$\tilde{J}(\delta x) = \frac{1}{2}\delta x^T B^{-1}\delta x + \frac{1}{2}(H\delta x - d)^T R^{-1}(H\delta x - d), \tag{25}$$

see (20), which is equivalent to solving

$$(B^{-1} + H^T R^{-1} H)\delta x = H^T R^{-1} d \quad \text{or} \quad \delta x = (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} d.$$

Here we have neglected all sub- and superscripts for simplicity. The minimization of the cost function in (25) is often referred to as the primal approach, as the minimization takes place in model space. Using the Sherman-Morrison-Woodbury formula we write the solution as

$$\delta x = BH^T(HBH^T + R)^{-1} d,$$

which can be obtained by solving the smaller system in observation space $(HBH^T + R)\lambda = d$ and setting $\delta x = BH^T\lambda$. This method is known as dual formulation of 3D-Var/4D-Var, or PSAS (Physical-space Statistical Analysis System) [48]. It is easy to see that the cost function for the dual formulation is given by

$$D(\lambda) = \frac{1}{2}\lambda^T(HBH^T + R)\lambda - \lambda^T d.$$

If the dimension of the observation space p is significantly smaller than that of the model state space n, the dual formulation can reduce both memory usage and computational cost compared to the primal approach. Preconditioned conjugate gradient methods for the dual approach were considered in [92,96] and convergence properties of the primal and dual approaches were investigated in [1,64]. Most recently, so called B-preconditioned minimization algorithms for variational data assimilation were introduced in the paper [99], which also contains a good literature review on the dual formulation and PSAS.

3.4 | Weak constraint 4D-Var and saddle point formulation of the inner iteration

Weak constraint 4D-Var requires minimization over all state variables within the assimilation window and is therefore more computationally expensive. The incremental approach [50] for the more general weak constraint 4D-Var cost function (8) can be formulated as follows. We approximate the 4D-Var cost function by a quadratic function of an increment

$$\delta x^{(\ell)} = x^{(\ell+1)} - x^{(\ell)}, \tag{26}$$

where $x^{(\ell)} = [(x_0^{(\ell)})^T, (x_1^{(\ell)})^T, \dots, (x_N^{(\ell)})^T]^T$ denotes the ℓth iterate of the Gauss-Newton algorithm applied to weak-constraint 4D-Var. This increment $\delta x^{(\ell)}$ is a solution to the minimization of the linearized cost function

$$\tilde{J}^{\ell}(\delta x^{(\ell)}) \quad = \frac{1}{2}\|\delta x_0^{(\ell)} - b_0^{(\ell)}\|^2_{B^{-1}} + \frac{1}{2}\sum_{i=0}^{N}\|d_i^{(\ell)} - H_i\delta x_i^{(\ell)}\|^2_{R_i^{-1}} + \frac{1}{2}\sum_{i=0}^{N-1}\|\delta x_{i+1}^{(\ell)} - M_i\delta x_i^{(\ell)} - c_{i+1}^{(\ell)}\|^2_{Q_{i+1}^{-1}}, \tag{27}$$

where M_i and H_i, are linearizations of $\mathcal{M}_i$ and $\mathcal{H}_i$ about the current state trajectory $x^{(\ell)}$, and $b_0^{(\ell)} = x_0^b - x_0^{(\ell)}$, $d_i^{(\ell)} = y_i - \mathcal{H}_i(x_i^{(\ell)})$, and $c_{i+1}^{(\ell)} = \mathcal{M}_i(x_i^{(\ell)}) - x_{i+1}^{(\ell)}$. Dropping the superscript for the ℓth iterate for simplicity, the linearized cost function

(27) can be written more concisely as

$$\tilde{J}(\delta x) = \frac{1}{2}\|\mathbf{L}\delta x - b\|^2_{\mathbf{D}^{-1}} + \frac{1}{2}\|d - \mathbf{H}\delta x\|^2_{\mathbf{R}^{-1}}, \tag{28}$$

where $\delta x = [\delta x_0^T, \delta x_1^T, \dots, \delta x_N^T]^T$ and $\mathbf{L}$ and $\mathbf{H}$ are matrices of size $(N+1)n \times (N+1)n$ and $(N+1)p \times (N+1)n$, respectively:

$$\mathbf{L} = \begin{bmatrix} I & & & \\ -M_0 & I & & \\ & \ddots & \ddots & \\ & & -M_{N-1} & I \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} H_0 & & & \\ & H_1 & & \\ & & \ddots & \\ & & & H_N, \end{bmatrix} \tag{29}$$

which can be thought of as all-at-once model and observation operators over the assimilation window. Here we assume $y_i \in \mathbb{R}^p$, but this can be generalized to $y_i \in \mathbb{R}^{p_i}$. We assume there is no correlation between the errors at each time steps, and hence the covariance matrices are block diagonal matrices

$$\mathbf{D} = \operatorname{diag}(B, Q_1, \dots, Q_N) \in \mathbb{R}^{(N+1)n \times (N+1)n}, \quad \text{and} \quad \mathbf{R} = \operatorname{diag}(R_0, R_1, \dots, R_N) \in \mathbb{R}^{(N+1)p \times (N+1)p}. \tag{30}$$

Moreover, the vectors b and d are given by

$$b = [b_0^T, c_1^T, \dots, c_N^T]^T \in \mathbb{R}^{(N+1)n}, \quad \text{and} \quad d = [d_0^T, d_1^T, \dots, d_N^T]^T \in \mathbb{R}^{(N+1)p}.$$

The system above can be written as a saddle point problem [72-74,77], a form that recently has seen a lot of interest for data assimilation problems, see also [80,97]. With new variables

$$\lambda = \mathbf{D}^{-1}(b - \mathbf{L}\delta x) \quad \text{and} \quad \mu = \mathbf{R}^{-1}(d - \mathbf{H}\delta x),$$

the gradient of the cost function (28) provides a constraint and altogether the coupled linear system

$$\begin{bmatrix} \mathbf{D} & 0 & \mathbf{L} \\ 0 & \mathbf{R} & \mathbf{H} \\ \mathbf{L}^T & \mathbf{H}^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \\ \delta x \end{bmatrix} = \begin{bmatrix} b \\ d \\ 0 \end{bmatrix}, \tag{31}$$

a very large, sparse, symmetric indefinite saddle point system needs to be solved at every inner iteration. A vast amount of literature on saddle point problems and their solution via Krylov methods and preconditioners is available, see for example [21] and references therein. For this particular saddle point problem low-rank limited memory preconditioners exploiting the structure of the saddle point problem were proposed and analyzed in [73] (see also [91]). In the work [80,97] the Kronecker structure of the saddle point problem was used in order to compute low-rank solutions to GMRES. The convergence of the saddle point formulation of weak constrained 4D-Var was reviewed in [90], and spectral estimates for the saddle point system were obtained in [58].

3.5 | Hybrid methods and inexact approaches

It has been observed that incremental weak constraint 4D-Var for minimizing the large scale cost function (8) may diverge, hence the authors in [131] add a regularization term and thereby replace the Gauss-Newton approach by the Levenberg-Marquardt method. In addition they use an ensemble Kalman smoother (see Section 4 for a discussion on ensemble methods) within the minimization process and thereby apply a hybrid approach. Such hybrid methods combining variational and sequential ensemble approaches are popular as ensemble methods are naturally parallelizable and do not require adjoint operators [45].

A parallel-in-time approach for solving the strong constraint 4D-Var optimization problem was proposed in [147].

When both the model operator and the gradient are not available exactly, inexact methods need to be used. In [18,23] a Levenberg-Marquardt method is proposed and investigated for dealing with inexact gradients and Jacobians.

4 | THE KALMAN FILTER AND LOW-RANK APPROXIMATIONS

The Kalman filter is impractical for large dimensional systems. It requires the storage and evolution of large covariance matrices $P_i^{A/F}$, which are both not feasible for very high dimensional systems (eg, in oceanography and numerical weather prediction, the state dimension is of the order of 10^7 and higher). The propagation of the error covariance in (13) requires a number of integrations of the forward model equal to the dimension of the system. Moreover matrix inversion within the Kalman gain computation (16) is expensive. Hence a range of approximate Kalman filters have been developed for large systems, either by using a simplified or reduced order model [46,60,71] to propagate the covariance matrices (ie, to propagate the error statistics) or by using a reduced state space or error space [36,146,172]. Many of the approaches are quite similar and most of them rely on low-rank approximations of the error covariance matrices. We discuss some methods below.

4.1 | Reduced-rank Kalman filters

The singular evolutive extended (SEEK) Kalman filter algorithm (see, eg, [27,129,146,153]) is one of the best known reduced rank square root (RRSQRT) filters. It is assumed that the covariance matrices P arising within the algorithm have low-rank form and can be written as $P = SS^T$, where $S \in \mathbb{R}^{n\times r}$ with $r \ll n$. The factorization can be obtained via a truncated eigenvalue decomposition, for example. The Kalman filter equations are then rewritten using the matrices S_i^F and S_i^A, the low rank approximations of the forecast error and analysis error covariance matrix, respectively. The equation for the Kalman gain (16) becomes

$$K_i = S_i^F (H_i S_i^F)^T (H_i S_i^F (H_i S_i^F)^T + R_i)^{-1},$$

or, using the Sherman-Morrison Woodbury identity

$$K_i = S_i^F [I_r + (H_i S_i^F)^T R_i^{-1} H_i S_i^F]^{-1} (H_i S_i^F)^T R_i^{-1}.$$

Note that often R_i is a (block) diagonal matrix, and the latter version of the low rank Kalman gain can be computed at lower cost in $r \ll n$ dimensions. The analysis increment is $x_i^A - x_i^F = K_i(y_i - H_i(x_i^F))$, and therefore a linear combination of the columns of S_i^F. Substituting the low rank Kalman gain into (17) yields $P_i^A = S_i^A (S_i^A)^T$ with

$$S_i^A = S_i^F [I_r + (H_i S_i^F)^T R_i^{-1} H_i S_i^F]^{-1/2},$$

where the inverse of the square root is taken in the lower dimensional space of dimension r.

For the forecast step, the propagation of the error covariance matrix is done via

$$P_{i+1}^F = \tilde{S}_{i+1}^F (\tilde{S}_{i+1}^F)^T + Q_i, \quad \text{where} \quad \tilde{S}_{i+1}^F = M_i S_i^A,$$

or, for nonlinear models, via the finite difference approximation

$$\{\tilde{S}_{i+1}^F\}_\ell = \mathcal{M}_i(x_i^A + \{S_i^A\}_\ell) - \mathcal{M}_i(x_i^A), \quad \ell = 1, \dots, r, \tag{32}$$

where $\{\cdot\}_\ell$ refers to the ℓth column. In order to write $P_{i+1}^F = S_{i+1}^F (S_{i+1}^F)^T$, and conserving the rank r some assumptions need to be made about Q_i (see [27,174] for details), otherwise a rank reduction, for example via computing an SVD, may be required at every step in order to keep the rank of the covariance matrices small.

Algorithm 4. SEEK filter

Input: error covariance matrices R_i and Q_i, routines to apply model and observation operators $\mathcal{M}_i$ and $\mathcal{H}_i$ and their linearizations M_i and H_i, respectively, and observations y_i for $i = 0, \dots, N$.
Initialize the system state x_0^F and the corresponding error covariance matrix in low-rank form $P_0^F = S_0^F (S_0^F)^T$.
for $i = 0, \dots, N$ **do**
Compute low rank Kalman gain $K_i = S_i^F [I_r + (H_i S_i^F)^T R_i^{-1} H_i S_i^F]^{-1} (H_i S_i^F)^T R_i^{-1}$.
Compute state estimate $x_i^A = x_i^F + K_i(y_i - \mathcal{H}_i(x_i^F))$.
Compute error covariance estimate $P_i^A = S_i^A (S_i^A)^T$, where $S_i^A = S_i^F [I_r + (H_i S_i^F)^T R_i^{-1} H_i S_i^F]^{-1/2}$.
Compute the forecast state $x_{i+1}^F = \mathcal{M}_i(x_i^A)$.
Compute the forecast error covariance $P_{i+1}^F = \tilde{S}_{i+1}^F (\tilde{S}_{i+1}^F)^T + Q_i$ where $\tilde{S}_{i+1}^F$ is given by (32).
end for

The SEEK filter, shown in Algorithm 4 is just one example of a reduced rank square root filter (RRSQRT), see [39,172]. Another example, shown in Algorithm 5 is a more general version of the SEEK filter and makes sure the rank of the error covariance matrix does not increase during the iteration. Again, we drop the time index, and assume a low rank approximation of the error covariance matrix $P = SS^T$ is possible, where $S \in \mathbb{R}^{n \times r}$ with $r \ll n$. The Kalman gain can then be written as

$$K_i = S_i^F (L_i^F)^T (L_i^F (L_i^F)^T + R_i)^{-1}, \quad \text{where} \quad L_i^F = H_i S_i^F,$$

and simple computations (again using the Sherman Morrison Woodbury identity), show that

$$P_i^A = S_i^F (I_r - (L_i^F)^T (L_i^F (L_i^F)^T + R_i)^{-1} L_i^F)(S_i^F)^T,$$

and using a matrix square root (of a smaller $r \times r$ matrix), one can write $S_i^A = S_i^F (I_r - (L_i^F)^T (L_i^F (L_i^F)^T + R_i)^{-1} L_i^F)^{\frac{1}{2}}$. Several algorithms are available for computing the matrix square root, for example, [62,105], but often a Cholesky factorization is used. After the analysis step the dimension of the system is reduced, keeping only $r - s$ eigenmodes of $(S_i^A)^T S_i^A$, where $s < r$. This avoids an increase of the rank of the covariance matrix when variability through the model error (with covariance matrix Q_i) is introduced. The details of the RRSQRT filter are given in Algorithm 5.

Algorithm 5. Reduced rank square root filter (RRSQRT)

Input: error covariance matrices R_i and Q_i, routines to apply model and observation operators $\mathcal{M}_i$ and $\mathcal{H}_i$ and their linearizations M_i and H_i, respectively, and observations y_i for $i = 0, \dots, N$.
Initialize the system state x_0^F and the corresponding error covariance matrix in low-rank form $P_0^F = S_0^F (S_0^F)^T$ and decompose $Q_i = T_i T_i^T$, a rank s factorization of Q_i, with $s > r$.
for $i = 0, \dots, N$ **do**
Compute low rank Kalman gain $K_i = S_i^F (L_i^F)^T (L_i^F (L_i^F)^T + R_i)^{-1}$, where $L_i^F = H_i S_i^F$.
Compute state estimate $x_i^A = x_i^F + K_i(y_i - \mathcal{H}_i(x_i^F))$.
Compute the low rank factor of error covariance $S_i^A = S_i^F (I_r - (L_i^F)^T (L_i^F (L_i^F)^T + R_i)^{-1} L_i^F)^{\frac{1}{2}}$.
Compute an eigenvalue decomposition (or low-rank factorization) $V \Lambda V^T = (S_i^A)^T S_i^A$.
Select largest $r - s$ eigenvalues and corresponding eigenvectors $\tilde{V} := V(:, 1 : r - s)$, set $\tilde{S}_i^A = S_i^A \tilde{V}$.
Compute the forecast state $x_{i+1}^F = \mathcal{M}_i(x_i^A)$.
Compute low rank factor $S_{i+1}^F = [M_i \tilde{S}_i^A, T_i]$ of the forecast error covariance $P_{i+1}^F = S_{i+1}^F (S_{i+1}^F)^T$.
end for

Reduced rank filters are cheaper to implement than the full filtering algorithm, by using low rank approximations of covariance matrices. Only r forward model integrations are necessary. However, they still use linearizations of the

nonlinear operators $\mathcal{M}_i$ and $\mathcal{H}_i$ in order to propagate the error covariance matrices. The ensemble Kalman filter overcomes this issue and utilizes the full nonlinear model to propagate the covariances.

4.2 | Ensemble Kalman filter

The need to propagate a probability distribution is an important feature of the ensemble Kalman filter (EnKF). It is also a major challenge as the propagation of large covariance matrices of size n is very expense. The ensemble Kalman filter is also a reduced rank method, as it requires the propagation and analysis of a small number of ensemble members. It was proposed in [69] (see also [24,70,106,165]) and is essentially a Monte Carlo implementation of the Bayesian update. There are plenty of variants of the EnKF, with the same idea behind all of them, the difference is in the implementation detail. We describe two versions briefly, the stochastic EnKF and the ensemble transform Kalman filter (ETKF), where, in the latter case, the linear algebra is performed in the ensemble subspace which is of much smaller dimension than the state space or the observation space.

Algorithm 6. Ensemble Kalman filter (EnKF)

Input: error covariance matrices R_i, routines to apply forward model and observation operators $\mathcal{M}_i$ and $\mathcal{H}_i$, respectively, and observations y_i for $i = 0, \dots, N$.

Initialize the ensemble states $x^F_{k,0}$ where $k = 1, \dots, r$ (via random perturbations from the initial conditions x^F_0, for example).

for $i = 0, \dots, N$ **do**

Perturb observations $y_{k,i} = y_i + v_k$, where $v_k \sim \mathcal{N}(0, R_i)$, $k = 1, \dots, r$.

Compute the ensemble means

$$\bar{x}^F_i = \frac{1}{r}\sum_{k=1}^{r} x^F_{k,i}, \quad \bar{v} = \frac{1}{r}\sum_{k=1}^{r} v_k, \quad \bar{y}^F_i = \frac{1}{r}\sum_{k=1}^{r} \mathcal{H}_i(x^F_{k,i})$$

Compute the rectangular normalized ensemble matrices

$$[X^F]_{k,i} = \frac{x^F_{k,i} - \bar{x}^F_i}{\sqrt{r-1}}, \quad [L^F]_{k,i} = \frac{\mathcal{H}_i(x^F_{k,i}) - \bar{y}^F_i - (v_k - \bar{v})}{\sqrt{r-1}}.$$

Compute the (approximate Kalman) gain: $K_i = X^F_i (L^F_i)^T (L^F_i (L^F_i)^T)^{-1}$.

Update the analysis ensemble

$$x^A_{k,i} = x^F_{k,i} + K_i(y_{k,i} - \mathcal{H}_i(x^F_{k,i})), \quad k = 1, \dots, r.$$

Compute the forecast ensemble $x^F_{k,i+1} = \mathcal{M}_i(x^A_{k,i}), \quad k = 1, \dots, r.$

end for

Suppose we have r ensemble members, or prior samples, $\{x^F_k\}_{k=1}^r$. Note that here the subscript k denotes the ensemble index, we neglect the time index in this explanation for simplicity. The forecast error covariance can be estimated using the empirical covariance

$$P^F = \frac{1}{r-1}\sum_{k=1}^{r}(x^F_k - \bar{x}^F)(x^F_k - \bar{x}^F)^T, \quad \text{where} \quad \bar{x}^F = \frac{1}{r}\sum_{k=1}^{r} x^F_k,$$

or

$$P^F = X^F (X^F)^T \quad \text{where} \quad [X^F]_k = \frac{x^F_k - \bar{x}^F}{\sqrt{r-1}}, \tag{33}$$

and $[X^F]_k$ the kth column of the $n \times r$ matrix X^F. Each of the ensemble members is then updated using (14): $x_k^A = x_k^F + K(y - \mathcal{H}(x_k^F))$ to obtain a posterior ensemble $[X^A]_k$, where

$$[X^A]_k = \frac{x_k^A - \bar{x}^A}{\sqrt{r-1}}, \quad \text{with} \quad \bar{x}^A = \frac{1}{r}\sum_{k=1}^{r} x_k^A.$$

When computing the sample posterior covariance $P^A = X^A(X^A)^T$ it turns out this is underestimated compared to the BLUE in (15). A way around this is to perturb the observation vector, $y_k = y + v_k$, where $v_k \sim \mathcal{N}(0, R)$ and set

$$[L^F]_k = \frac{Hx_k^F - H\bar{x}^F - (v_k - \bar{v})}{\sqrt{r-1}}.$$

Then it can be shown [8] that $[X^A]_k = [X^F]_k - K[L^F]_k$ results in the correctly estimated $P^A = X^A(X^A)^T$. The Kalman gain can be computed using $K = X^F(L^F)^T(L^F(L^F)^T)^{-1}$. Moreover it can be shown that within the algorithm we only require the application of the nonlinear operator $\mathcal{H}$ to the ensemble members, rather than its linearized version H.

The full version of the (stochastic) EnKF is given in Algorithm 6. Note that k is the ensemble index, i is, as before, the time index. For more details on the algorithm we refer to [8,103]. An error analysis for the ensemble Kalman filter analysis step was performed in [116].

Algorithm 7. Ensemble Transform Kalman filter (ETKF)

Input: error covariance matrices R_i, routines to apply forward model and observation operators $\mathcal{M}_i$ and $\mathcal{H}_i$, respectively, and observations y_i for $i = 0, \dots, N$. $V \in \mathbb{R}^{r\times r}$ an orthogonal matrix such that $V\mathbf{1} = \mathbf{1}$.
Initialize the ensemble states $\{x_k^F\}_{k=1,\dots,r}$ at time $i = 0$ and set $E^F = [x_1^F, \dots, x_r^F]$.
for $i = 0, \dots, N$ **do**
  Compute ensemble mean $\bar{x}^F = E^F\mathbf{1}/r$ and ensemble matrix $X^F = (E^F - \bar{x}^F\mathbf{1}^T)/\sqrt{r-1}$.
  Compute the observation mean $\bar{y} = Y\mathbf{1}/r$ where $Y = \mathcal{H}_i(E^F)$.
  Compute normalized observation ensemble $\tilde{L}^F = R_i^{-\frac{1}{2}}(Y - \bar{y}\mathbf{1}^T)/\sqrt{r-1}$.
  Compute normalized innovation vector $d = R_i^{-\frac{1}{2}}(y_i - \bar{y})$ and set $W = (I_r + (\tilde{L}^F)^T\tilde{L}^F)^{-1}$.
  Compute ensemble space coefficient vector $w^A = W(\tilde{L}^F)^T d$.
  Update state estimate ensemble $E^A = \bar{x}^F\mathbf{1}^T + X^F(w^A\mathbf{1}^T + \sqrt{r-1}W^{\frac{1}{2}}V)$.
  Compute the forecast ensemble $E^F = \mathcal{M}_i(E^A)$.
end for

For the deterministic version of the ensemble Kalman filter [24], the ETKF which is illustrated in Algorithm 7, we again write the forecast error covariance in low rank form (33), with ensembles $\{x_k^F\}_{k=1}^r$. It is then assumed that the state estimate x^A is of the form $x^A = \bar{x}^F + X^F w^A$, where w^A is a vector of coefficients in the small dimensional ensemble subspace $\mathbb{R}^r$, and $X^F \in \mathbb{R}^{n\times r}$. Using (14), the mean of the analysis vector is given by $x^A = \bar{x}^F + K(y - \mathcal{H}(\bar{x}^F))$. Hence, we obtain

$$\bar{x}^F + X^F w^A = \bar{x}^F + K(y - \mathcal{H}(\bar{x}^F)),$$

and with the low rank approximation $P^F = X^F(X^F)^T$ within the Kalman gain K, and using the Sherman-Morrison-Woodbury formula again, we derive

$$w^A = (I_r + (L^F)^T R^{-1} L^F)^{-1}(L^F)^T R^{-1}(y - \mathcal{H}(\bar{x}^F)), \quad \text{where} \quad L^F = \mathcal{H}X^F,$$

and hence the Kalman gain is computed in the low dimensional ensemble space. For computing the posterior ensemble covariance matrix we proceed as in the RRSQRT derivation. This yields

$$X^A = X^F(I_r + (L^F)^T R^{-1} L^F)^{-\frac{1}{2}} V,$$

where V is an arbitrary orthogonal matrix. To ensure that $X^A\mathbf{1}=0$, that is, the updated perturbations are centered at x^A (similar to $X^F\mathbf{1}=0$) it is sufficient for $V\mathbf{1}=\mathbf{1}$ to hold. Here $\mathbf{1}$ is the vector of all ones. The posterior ensemble is then given by

$$x_k^A = x^A + \sqrt{r-1}X^F\left[W^{\frac{1}{2}}V\right]_k = \bar{x}^F + X^F\left(w^A + \sqrt{r-1}\left[W^{\frac{1}{2}}V\right]_k\right),$$

where $W=(I_r+(L^F)^TR^{-1}L^F)^{-1}$. Note that within the ETKF, all matrix inversions and matrix square roots are carried out in the lower dimensional ensemble subspace of dimension r. Moreover, with a small number of ensemble members, only r applications of the expensive forward model $\mathcal{M}_i$ and the operator $\mathcal{H}_i$ are necessary.

Using RRSQRT filters and EnKFs results in the increments being confined in a subspace spanned by the columns of the low rank matrix . There are localization methods which overcome these issues, for details we refer to [8,108,142] and references therein. A whole range of variations of the ensemble and square root filters have been developed over the years, see the references in [8,137].

4.3 | Iterative solvers within the Kalman filter

As mentioned at the end of Section 2, a variational form of the Kalman filter can be formulated and state estimate and posterior covariance are given by minimum and inverse Hessian of a quadratic cost function, respectively [13]. When the preconditioned CG method is used, this leads to a conjugate gradient ensemble Kalman filter, which has been discussed in [15]. Low-rank approximations to covariance and inverse covariance matrices can be obtained by exploiting the connection between conjugate gradient and Lanczos iterations [14]. Moreover, in a similar way, the use of limited memory BFGS [141] within the Kalman filter estimate has been investigated in [10].

5 | MODEL REDUCTION AND DIMENSION REDUCTION APPROACHES

Data assimilation problems are often large, in particular when the model is described by discretized time-dependent partial differential equations and when large amounts of data need to be assimilated. This leads to either a large optimization problem for variational data assimilation, or, the solution to large linear systems, eventually, for the optimization problem and the Kalman filter. Over the years several reduction techniques have been proposed for data assimilation problems (or, inverse problems) which we will describe here briefly.

We will collect results from two different ideas. Usually the state space dimension n in data assimilation is very large. This leads to very large covariance matrices which need to be stored, manipulated and inverted, which is expensive. Therefore, one approach is the reduction of the dimension of these covariance matrices, usually by using low-rank approximations.

A second approach considers the expensive, PDE-based model operator $\mathcal{M}$ (or $\mathcal{M}_i$, $i=0,\ \dots\ ,N$). In order to solve the optimization problem in 4D-Var, or, in order to apply sampling based methods such as the ensemble Kalman filter (or, more general Markov Chain Monte Carlo methods), this model has to be evaluated many times at several points in space, which is expensive. Therefore reducing the dimension of that model operator is a second way of applying dimension reduction.

5.1 | Reduced rank methods—Reducing the dimension of the covariance matrix

A key idea in dimension reduction for Bayesian inference is to exploit the fact that in updating the prior to the posterior, some directions in the high-dimensional state space are more important than others. In the case of Gaussian posteriors, this fact is quantified by the rate of decay of the eigenvalues of the Hessian of the data misfit [31,78]. The decay leads to a low-rank approximation of the Hessian (and hence the posterior covariance). Quantitative error analysis of approximation methods for the posterior distribution in Bayesian inference have been performed in [160].

We have already explained several reduced rank approaches applied to the Kalman filter in Section 4, which are based on reduced rank covariance matrices, so we will not repeat details here.

In recent years, a whole range of hybrid methods were developed, which combine ensemble (and hence low-rank) Kalman filters with variational data assimilation. This can be done in several ways and we point to [11] for a review of those ideas. One such method (often referred to EnVar) uses the low rank covariance matrix that arises within the EnKF, and applies it within variational data assimilation as the background error covariance matrix, $B = SS^T$, where $S \in \mathbb{R}^{n\times r}$ and $r \ll n$. The resulting system has the same form as preconditioned 4D-Var, however with a low-rank version of the Hessian, and hence the optimization problem is solved in the reduced system dimension, see [93].

As discussed above, it is known that for linear inverse problems, the inverse of the Hessian is an approximation of the posterior error covariance. From a control theoretic view point, there is an equivalence between the Hessian and the observability Gramian. Hessian based model reduction using this viewpoint was investigated in [16,17,128]

In [20], the Kronecker product structure of a discretized linear partial differential operator and a low-rank Arnoldi method were used to approximate posterior error covariance matrices of Gaussian posteriors using the idea of low-rank Hessians [31,78,160].

5.2 | Model order reduction applied to the forward model operator

A second approach for reducing the cost within sequential and variational data assimilation is reducing the cost of the forward and adjoint models by applying reduced order models (ROMs). The aim of model order reduction (MOR) is to find models that approximate and reflect the dynamics of the underlying large-scale system accurately, in ways that enable the reduction process to be implemented efficiently. There is a large number of MOR techniques available, and many of them have been very popular in the system theoretic community [19].

The earliest reduction approaches for 4D-Var suggest using a simplified operator or a coarser grid within the minimization of incremental 4D-Var [109,166].

In [71] balanced truncation [132], a control theoretic approach for MOR, is applied to the Kalman filter. The linearized model and observation operators $M_i \in \mathbb{R}^{n\times n}$ and $H_i \in \mathbb{R}^{p\times n}$ are projected onto lower dimensional subspaces,

$$\hat{M}_i = U^T M_i V \in \mathbb{R}^{r\times r}, \quad \hat{H}_i = H_i V \in \mathbb{R}^{p\times r}, \tag{34}$$

where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{n\times r}$ are projection matrices satisfying $U^T V = I$, obtained through balanced truncation. Balanced truncation is a method that retains the dominant observable and reachable states, which are the important ones for the system dynamics (after transforming both state and observation equation so that reachable and observable states can be expressed in the same basis). In [71] it is assumed that the time-dependent system underlying the problem has a time-invariant dominant part on which balanced truncation is performed.

MOR via balanced truncation was also proposed for incremental 4D-Var in [25,26,124,125]. Model and observation operators are projected as in (34), δx is restricted to $\delta\hat{x} = U^T\delta x \in \mathbb{R}^r$, $r \ll n$, and the background error covariance matrix, B, is projected onto $U^T BU$. The approach was extended to weak constraint 4D-Var in [81].

In [63,152] it is assumed that the initial state x_0 in 4D-Var is contained in a space of reduced dimension $r \ll n$ about the background state,

$$x_0 = x_0^B + \sum_{i=1}^{r} c_i w_i, \quad r \ll n,$$

where c_i are real coefficients and w_i are linearly independent vectors containing the variability in the system. Minimization of the reduced cost function then takes place in the reduced space of dimension r. The authors in [7] use a truncated SVD approach to solve the 4D-Var problem in reduced spaces.

Proper orthogonal decomposition (POD) methods were applied to variational data assimilation by several authors [37,55,61,173]. Snapshots are taken at various time steps from the model trajectory. An SVD is then taken of the matrix of snapshots $X = [x_1, \ldots, x_r]$, that is, $X = U\Sigma V^T$. Finally, x_0 is then projected onto the POD space spanned by the r most important left singular vectors, that is, the ones corresponding to the largest singular values. The number of those left singular vectors is chosen significantly smaller than the dimension of the state space. POD was applied in [3] to the adjoint in order to compute a reduced model. The work in [161] refines this work on POD by considering the discrete empirical interpolation method (DEIM) within POD for nonlinear dynamical systems.

A challenge for applying MOR techniques within data assimilation is the nonlinearity and time-dependence of the forward model operator. One idea is to use a data driven and online approach applied to the ROMs, where updates to the basis vectors of the ROMs are computed from combinations of reduced solutions, snapshots and adjoint information [144].

Dimension reduction is an important tool for Bayesian inverse problems as one often has to perform Markov chain Monte Carlo (MCMC) sampling to access the posterior distribution. However, each MCMC sample requires an expensive forward model solve. In [51-53], methods for dimension reduction for nonlinear Bayesian inverse problems were described. The goal of these method is to approximate the likelihood using a reduced model. In [67], POD-DEIM is successfully applied to the likelihood function in order to reduce the cost of each MCMC draw. A review of multi-fidelity methods was recently published [145]. For additional ideas for dimension reduction methods within data assimilation we refer to [8] and references therein.

6 | BAYESIAN INFERENCE AND TIKHONOV REGULARIZATION, AND OTHER ASPECTS WITHIN DATA ASSIMILATION

One aspect we have not considered in detail in this article is the link between data assimilation and Tikhonov regularization. We refer to [82] and references therein. A key task in the Bayesian inverse problem framework is finding an informative prior. This corresponds to finding a computationally efficient penalty term in the regularization approach, see [33,35]. Most of the work in this area considers Gaussian priors and linear, but very large scale static inverse problems. In reference [41] methods based on Golub-Kahan bidiagonalization for computing solutions based on Tikhonov regularization are used, which avoid computing inverses and square roots of large covariance matrices. The idea was generalized to dynamic inverse problems, which also allows the use of a wide class of spatio-temporal priors [42,154].

When designing an observation or sensor network for data assimilation another important aspect not considered here is the placement of sensors. An important tool for this is computing the so-called observation impact, which provides a measure for important or redundant information. Mathematically this results in analyzing the sensitivity of states with respect to the data, and eventually in a large linear system to solve. For more information and solution methods, including low-rank approaches, we refer to [43,44,169] and references therein.

Data assimilation using different regularization terms was considered in [30,148], efficient solution techniques for such approaches require ideas from numerical linear algebra.

There is a whole range of ideas for data assimilation we have not considered here. For literature on particle filters for Bayesian inference, for example, we refer to [8,40] and references therein.

7 | SOFTWARE

Software development for data assimilation algorithms is very much driven by geoscientists. Several packages are available to download and test algorithms, these also often provide examples. We list a few here.

PDAF, a software environment (written in FORTRAN) for ensemble data assimilation was developed by researchers from the Alfred Wegener Institute for Polar and Marine Research and is freely available [136]. DATes is a recently developed data assimilation testing suite written in PYTHON [9]. Finally, DART [5] is a FORTRAN software package developed and maintained at the National Center for Atmospheric Research.

MATLAB code, which is applied to a range of simple examples, both for Kalman filters and variational data assimilation, is available in the book by Law et al [119]. Further MATLAB codes are provided which accompany the book [150].

8 | CONCLUSIONS

In the era of "big data" it is important to be able to process and evaluate data, gain insight from data, extract knowledge from data and make predictions from data. However, making important decisions based only on data can be misleading as they might be incomplete or erroneous. Therefore, it is often crucial to combine data with mathematical models. This approach leads to the important area of *data assimilation* which is evolving rapidly and continuously.

Data assimilation uses tools from many different fields of mathematics, such as statistics, optimization, machine learning, numerical linear algebra, mathematical modeling and scientific computing, to name a few.

For example, classical optimization tools in variational data assimilation are constantly adapted to faster algorithms and new supercomputers. Variational data assimilation typically only provides a point estimate, but no uncertainty quantification. However, approximations of a-posteriori covariance matrices can be computed using efficient numerical linear algebra techniques.

On the other hand, statistical learning techniques, such as Kalman filters and ensemble methods are regularly seeing improvements. Those techniques do provide uncertainty quantification as they are merely Monte Carlo implementation of the Bayesian update.

We have seen that in the field of data assimilation, traditional computational scientific modeling meets the area of statistics and data science in order to produce new and better algorithms and methods.

In this review we have stated and explained the most established data assimilation methods, starting from the point of view of Bayesian inference. We have focused on problems arising within the numerical linear algebra for these methods, that is, the solution to linear systems, preconditioning, and matrix methods for filtering and model reduction, and thereby given a extended (but by no means complete) review of the existing literature in numerical methods for data assimilation.

With the increasing size and complexity of datasets and larger models after discretization of partial differential equations, many data assimilation problems involve operations on matrices with millions or billions of elements. This amount of large matrices brings new computational challenges to classical numerical linear algebra algorithms, for example solving large systems of linear equations, large linear (and nonlinear) regression problems, constructing low-rank matrix approximation, etc.

The efficiency of these linear algebra applications is essential for the performance of data assimilation methods.

Hence, this review shows that efficient numerical linear algebra techniques and matrix computation tools are crucial within the subject of data driven modeling and data assimilation, even more so when the data and models become even larger.

ACKNOWLEDGEMENTS

The research of the author has been partially funded by Deutsche Forschungsgemeinschaft (DFG)—SFB1294/1—318763901 (associated member). Open access funding enabled and organized by Projekt DEAL.

REFERENCES

[1] A. E. Akkraoui and P. Gauthier, *Convergence properties of the primal and dual forms of variational data assimilation*, Q. J. R. Meteorol. Soc. **136** (2010), 107–115.

[2] M. Allmaras et al., *Estimating parameters in physical models through Bayesian inversion: A complete example*, SIAM Rev. **55** (2013), 149–167.

[3] M. U. Altaf et al., *A reduced adjoint approach to variational data assimilation*, Comput. Methods Appl. Mech. Engrg. **254** (2013), 1–13.

[4] E. Anderson et al., *Diagnosis of background errors for radiances and other observable quantities in a variational data assimilation scheme, and the explanation of a case of poor convergence*, Q. J. R. Meteorol. Soc. **126** (2000), 1455–1472.

[5] J. Anderson et al., *The data assimilation research testbed: A community facility*, Bull. Am. Meteorol. Soc. **90** (2009), 1283–1296.

[6] A. Apte et al., *Data assimilation: Mathematical and statistical perspectives*, Internat. J. Numer. Methods Fluids **56** (2008), 1033–1046.

[7] R. Arcucci et al., *Optimal reduced space for variational data assimilation*, J. Comput. Phys. **379** (2019), 51–69.

[8] M. Asch, M. Bocquet, and M. Nodet, *Data assimilation: Methods, algorithms and applications*, SIAM, New York, 2016.

[9] A. Attia and A. Sandu, *Dates: A highly-extensible data assimilation testing suite v1. 0*, arXiv preprint arXiv:1704.05594, 2017.

[10] H. Auvinen et al., *Large-scale Kalman filtering using the limited memory BFGS method*, Electron. Trans. Numer. Anal. **35** (2009), 217–233.

[11] R. N. Bannister, *A review of operational methods of variational and ensemble-variational data assimilation*, Q. J. R. Meteorol. Soc. **143** (2017), 607–633.

[12] J. M. Bardsley, *Computational uncertainty quantification for inverse problems*, in *Computational Science and Engineering*, SIAM, New York, 2018.

[13] J. M. Bardsley, *A matrix theoretic derivation of the Kalman filter*, in *2017 MATRIX Annals*, Springer, Berlin, 2019, 505–513.

[14] J. M. Bardsley et al., *Krylov space approximate Kalman filtering*, Numer. Linear Algebra Appl. **20** (2013), 171–184.

[15] J. M. Bardsley et al., *An ensemble Kalman filter using the conjugate gradient sampler*, Int. J. Uncertain. Quantif. **3** (2013), 357–370.

[16] O. Bashir et al., *Hessian-based model reduction for large-scale data assimilation problems*, in *International Conference on Computational Science*, Springer, Berlin, 2007, 1010–1017.

[17] O. Bashir et al., *Hessian-based model reduction for large-scale systems with initial-condition inputs*, Internat. J. Numer. Methods Engrg. **73** (2008), 844–868.

[18] S. Bellavia, S. Gratton, and E. Riccietti, *A Levenberg-Marquardt method for large nonlinear least-squares problems with dynamic accuracy in functions and gradients*, Numer. Math. **140** (2018), 791–825.

[19] P. Benner et al., *Model reduction and approximation: Theory and algorithms*, in *Computational Science and Engineering*, Vol **15**, SIAM, New York, 2017.

[20] P. Benner, Y. Qiu, and M. Stoll, *Low-rank eigenvector compression of posterior covariance matrices for linear Gaussian inverse problems*, SIAM/ASA J. Uncertain. Quantif. **6** (2018), 965–989.

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

[22] D. Béréziat and I. Herlin, *Solving ill-posed image processing problems using data assimilation*, Numer. Algorithms **56** (2011), 219–252.

[23] E. Bergou, S. Gratton, and L. N. Vicente, *Levenberg-Marquardt methods based on probabilistic gradient models and inexact subproblem solution, with application to data assimilation*, SIAM/ASA J. Uncertain. Quantif. **4** (2016), 924–951.

[24] C. H. Bishop, B. J. Etherton, and S. J. Majumdar, *Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects*, Mon. Weather Rev. **129** (2001), 420–436.

[25] C. Boess, *Using model reduction techniques within the Incremental 4D-Var method*, PhD thesis, University of Bremen, 2008.

[26] C. Boess et al., *State estimation using model order reduction for unstable systems*, Comput. Fluids **46** (2011), 155–160.

[27] P. Brasseur and J. Verron, *The SEEK filter method for data assimilation in oceanography: A synthesis*, Ocean Dyn. **56** (2006), 650–661.

[28] K. L. Brown, I. Y. Gejadze, and A. Ramage, *A multilevel approach for computing the limited-memory Hessian and its inverse in variational data assimilation*, SIAM J. Sci. Comput. **38** (2016), A2934–A2963.

[29] S. D. Brown, *The Kalman filter in analytical chemistry*, Anal. Chim. Acta **181** (1986), 1–26.

[30] C. J. Budd, M. A. Freitag, and N. K. Nichols, *Regularization techniques for ill-posed inverse problems in data assimilation*, Comput. Fluids **46** (2011), 168–173.

[31] T. Bui-Thanh et al., *Extreme-scale uq for Bayesian inverse problems governed by pdes*, in *Proceedings of the international conference on high performance computing, networking, storage and analysis*, IEEE Computer Society Press, New York, 2012, 3.

[32] M. D. Butala et al., *Tomographic imaging of dynamic objects with the ensemble Kalman filter*, IEEE Trans. Image Process. **18** (2009), 1573–1587.

[33] D. Calvetti et al., *Bayes meets Krylov: Statistically inspired preconditioners for CGLS*, SIAM Rev. **60** (2018), 429–461.

[34] D. Calvetti and E. Somersalo, *An introduction to Bayesian scientific computing: Ten lectures on subjective computing*, Vol **2**, Springer Science & Business Media, Berlin, 2007.

[35] D. Calvetti and E. Somersalo, *Inverse problems: From regularization to Bayesian inference*, Wiley Interdiscip. Rev. Comput. Stat. **10** (2018), e1427.

[36] M. A. Cane et al., *Mapping tropical pacific sea level: Data assimilation via a reduced state space Kalman filter*, J. Geophys. Res. Oceans **101** (1996), 22599–22617.

[37] Y. Cao et al., *A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition*, Internat. J. Numer. Methods Fluids **53** (2007), 1571–1583.

[38] A. Carrassi et al., *Data assimilation in the geosciences: An overview of methods, issues, and perspectives*, Wiley Interdiscip. Rev. Clim. Change **9** (2018), e535.

[39] J. Chandrasekar et al., *Cholesky-based reduced-rank square-root Kalman filtering*, in *2008 American Control Conference*, Seattle, WA, IEEE 2008, 3987–3992.

[40] Z. Chen, *Bayesian filtering: From Kalman filters to particle filters, and beyond*, Statistics **182** (2003), 1–69.

[41] J. Chung and A. K. Saibaba, *Generalized hybrid iterative methods for large-scale Bayesian inverse problems*, SIAM J. Sci. Comput. **39** (2017), S24–S46.

[42] J. Chung et al., *Efficient generalized Golub-Kahan based methods for dynamic inverse problems*, Inverse Problems **34** (2018), 024005.

[43] A. Cioaca and A. Sandu, *An optimization framework to improve 4d-var data assimilation system performance*, J. Comput. Phys. **275** (2014), 377–389.

[44] A. Cioaca, A. Sandu, and E. de Sturler, *Efficient methods for computing observation impact in 4d-var data assimilation*, Comput. Geosci. **17** (2013), 975–990.

[45] A. M. Clayton, A. C. Lorenc, and D. M. Barker, *Operational implementation of a hybrid ensemble/4d-var global data assimilation system at the Met Office*, Q. J. R. Meteorol. Soc. **139** (2013), 1445–1461.

[46] S. E. Cohn and R. Todling, *Approximate data assimilation schemes for stable and unstable dynamics*, J. Meteor. Soc. Japan Ser. II **74** (1996), 63–75.

[47] A. Corigliano and S. Mariani, *Parameter identification in explicit structural dynamics: Performance of the extended Kalman filter*, Comput. Methods Appl. Mech. Engrg. **193** (2004), 3807–3835.

[48] P. Courtier, *Dual formulation of four-dimensional variational assimilation*, Q. J. R. Meteorol. Soc. **123** (1997), 2449–2461.

[49] P. Courtier et al., *Important literature on the use of adjoint, variational methods and the Kalman filter in meteorology*, Tellus A **45** (1993), 342–357.

[50] P. Courtier, J.-N. Thépaut, and A. Hollingsworth, *A strategy for operational implementation of 4D-Var, using an incremental approach*, Q. J. R. Meteorol. Soc. **120** (1994), 1367–1387.

[51] T. Cui et al., *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems **30** (2014), 114015.

[52] T. Cui, Y. Marzouk, and K. E. Willcox, *Data-driven model reduction for the Bayesian solution of inverse problems*, Int. J. Numer. Methods Engrg **102** (2015), 966–990.

[53] T. Cui, Y. Marzouk, and K. E. Willcox, *Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction*, J. Comp. Phys. **315** (2016), 363–387.

[54] J. K. Cullum and R. A. Willoughby, *Lanczos algorithms for large symmetric eigenvalue computations*, in *Theory*, Vol **1**, **41**, Philadelphia: SIAM, 2002.

[55] D. N. Daescu and I. M. Navon, *Efficiency of a POD-based reduced second-order adjoint model in 4D-Var data assimilation*, Internat. J. Numer. Methods Fluids **53** (2007), 985–1004.

[56] R. Daley, *Atmospheric data analysis*, Cambridge University Press, Cambridge, MA, 1993.

[57] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, in *Handbook of Uncertainty Quantification*, Switzerland: Springer International Publishing, 2016, 1–118.

[58] I. Daužickaitė et al., Spectral estimates for saddle point matrices arising in weak constraint four-dimensional variational data assimilation, 2019. arXiv:1908.07949.

[59] L. Debreu et al., *Multigrid solvers and multigrid preconditioners for the solution of variational data assimilation problems*, Q. J. R. Meteorol. Soc. **142** (2015), 515–528.

[60] D. P. Dee, *Simplification of the Kalman filter for meteorological data assimilation*, Q. J. R. Meteorol. Soc. **117** (1991), 365–384.

[61] G. Dimitriu, N. Apreutesei, and R. Ştefănescu, *Numerical simulations with data assimilation using an adaptive pod procedure*, in *International Conference on Large-Scale Scientific Computing*, Springer, Berlin, 2009, 165–172.

[62] V. Druskin and L. Knizhnerman, *Extended krylov subspaces: Approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl. **19** (1998), 755–771.

[63] S. Durbiano, *Vecteurs caractéristiques de modèles océaniques pour la réduction d'ordre en assimilation de données*, PhD thesis, Université Joseph Fourier Grenoble, 2001.

[64] A. El Akkraoui et al., *Intercomparison of the primal and dual formulations of variational data assimilation*, Q. J. R. Meteorol. Soc. **134** (2008), 1015–1025.

[65] A. El Akkraoui, Y. Trémolet, and R. Todling, *Preconditioning of variational data assimilation and the use of a bi-conjugate gradient method*, Q. J. R. Meteorol. Soc. **139** (2013), 731–741.

[66] H. Elbern, H. Schmidt, and A. Ebel, *Variational data assimilation for tropospheric chemistry modeling*, J. Geophys. Res. Atmos. **102** (1997), 15967–15985.

[67] H. C. Elman and A. Onwunta, *Reduced-order modeling for nonlinear Bayesian statistical inverse problems*, 2019. arXiv:1909.02539.

[68] R. Engbert et al., *Sequential data assimilation of the stochastic seir epidemic model for regional covid-19 dynamics*, medRxiv (2020), 1–19. Doi: 10.1101/2020.04.13.20063768.

[69] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, J. Geophys. Res. **99** (1994), 10143–10162.

[70] G. Evensen, *Data assimilation: The ensemble Kalman filter*, Springer Science & Business Media, Berlin, 2009.

[71] B. F. Farrell and P. J. Ioannou, *State estimation using a reduced-order Kalman filter*, J. Atmos. Sci. **58** (2001), 3666–3680.

[72] M. Fisher and H. Auvinen, *Long window 4D-Var*, in *Proc. Seminar on Data Assimilation for Atmosphere and Ocean*, Shinfield Park, Readingr: ECMWF, 2011, 189–202.

[73] M. Fisher et al., *Low rank updates in preconditioning the saddle point systems arising from data assimilation problems*, Optim. Method. Softw. (2016), 33(1), 1–25.

[74] M. Fisher and S. Gürol, *Parallelisation in the time dimension of four-dimensional variational data assimilation*, Q. J. R. Meteorol. Soc. (2017), 143(703), 1136–1147.

[75] M. Fisher, M. Leutbecher, and G. A. Kelly, *On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation*, Q. J. R. Meteorol. Soc. **131** (2005), 3235–3246.

[76] M. Fisher et al., *Data assimilation in weather forecasting: A case study in pde-constrained optimization*, Optim. Eng. **10** (2009), 409–426.

[77] M. Fisher, Y. Trémolet, H. Auvinen, D. Tan, and P. Poli, *Weak-constraint and long-window 4D-Var*, Tech. Rep. 655, ECMWF, 2011.

[78] H. P. Flath et al., *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM J. Sci. Comput. **33** (2011), 407–432.

[79] S. J. Fletcher, *Data assimilation for the geosciences: From theory to application*, Elsevier, Amsterdam, 2017.

[80] M. A. Freitag and D. L. H. Green, *A low-rank approach to the solution of weak constraint variational data assimilation problems*, J. Comput. Phys. **357** (2018), 263–281.

[81] M. A. Freitag and D. L. H. Green, *Projection methods for weak constraint variational data assimilation*, submitted, 2019.

[82] M. A. Freitag and R. W. E. Potthast, *Synergy of inverse problems and data assimilation techniques*, in *Large scale inverse problems, of Radon Ser. Comput. Appl. Math.*, Vol **13**, De Gruyter, Berlin, 2013, 1–53.

[83] I. Y. Gejadze, F.-X. Le Dimet, and V. Shutyaev, *On analysis error covariances in variational data assimilation*, SIAM J. Sci. Comput. **30** (2008), 1847–1874.

[84] I. Y. Gejadze, V. P. Shutyaev, and F.-X. Le Dimet, *Hessian-based covariance approximations in variational data assimilation*, Russ. J. Numer. Anal. M. **33** (2018), 25–39.

[85] A. Gelb, *Applied optimal estimation*, Cambridge, MA: MIT Press, 1974.

[86] M. Ghil, *Meteorological data assimilation for oceanographers. Part I: Description and theoretical framework*, Dynam. Atmos. Oceans **13** (1989), 171–218.

[87] M. Ghil and P. Malanotte-Rizzoli, *Data assimilation in meteorology and oceanography*, in *Advances in Geophysics*, Vol **33**, Elsevier, Amsterdam, 1991, 141–266.

[88] P. E. Gill, W. Murray, and M. H. Wright, *Practical optimization*, SIAM, New York, 2019.

[89] G. H. Golub and C. F. Van Loan, *Matrix computations*, Vol **3**, Baltimore, Maryland: Johns Hopkins University Press, 2012.

[90] S. Gratton et al., *Guaranteeing the convergence of the saddle formulation for weakly constrained 4D-Var data assimilation*, Q. J. R. Meteorol. Soc. **144** (2018), 2592–2602.

[91] S. Gratton et al., *A note on preconditioning weighted linear least-squares, with consequences for weakly constrained variational data assimilation*, Q. J. R. Meteorol. Soc. **144** (2018), 934–940.

[92] S. Gratton, S. Gürol, and P. L. Toint, *Preconditioning and globalizing conjugate gradients in dual space for quadratically penalized nonlinear-least squares problems*, Comput. Optim. Appl. **54** (2013), 1–25.

[93] S. Gratton et al., *A reduced and limited-memory preconditioned approach for the 4D-var data-assimilation problem*, Q. J. R. Meteorol. Soc. **137** (2011), 452–466.

[94] S. Gratton, A. S. Lawless, and N. K. Nichols, *Approximate Gauss–Newton methods for nonlinear least squares problems*, SIAM J. Optim. **18** (2007), 106–132.

[95] S. Gratton, P. L. Toint, and J. Tshimanga, *Conjugate gradients versus multigrid solvers for diffusion-based correlation models in data assimilation*, Q. J. R. Meteorol. Soc. **139** (2013), 1481–1487.

[96] S. Gratton and J. Tshimanga, *An observation-space formulation of variational assimilation using a restricted preconditioned conjugate gradient algorithm*, Q. J. R. Meteorol. Soc. **135** (2009), 1573–1585.

[97] D. L. H. Green, *Model order reduction for large-scale data assimilation problems*, PhD thesis, University of Bath, 2019.

[98] A. K. Griffith, *Data assimilation for numerical weather prediction using control theory*, PhD thesis, Department of Mathematics, 1997.

[99] S. Gürol et al., *B-preconditioned minimization algorithms for variational data assimilation with the dual formulation*, Q. J. R. Meteorol. Soc. **140** (2014), 539–556.

[100] F. Gustafsson and G. Hendeby, *Some relations between extended and unscented Kalman filters*, IEEE Trans. Signal Process. **60** (2011), 545–555.

[101] S. A. Haben, A. S. Lawless, and N. K. Nichols, *Conditioning and preconditioning of the variational data assimilation problem*, Comput. Fluids **46** (2011), 252–256.

[102] P. C. Hansen, *Discrete inverse problems: Insight and algorithms*, Vol **7**, SIAM, New York, 2010.

[103] J. Harlim, *Data-driven computational methods: Parameter and operator estimations*, Cambridge University Press, Cambridge, MA, 2018.

[104] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Vol **49**, NBS, Washington, DC, 1952.

[105] N. J. Higham, *Stable iterations for the matrix square root*, Numer. Algorithms **15** (1997), 227–242.

[106] P. L. Houtekamer and H. L. Mitchell, *Data assimilation using an ensemble Kalman filter technique*, Mon. Weather Rev. **126** (1998), 796–811.

[107] J. Humpherys, P. Redd, and J. West, *A fresh look at the Kalman filter*, SIAM Rev. **54** (2012), 801–823.

[108] B. R. Hunt, E. J. Kostelich, and I. Szunyogh, *Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter*, Phys. D **230** (2007), 112–126.

[109] K. Ide et al., *Unified notation for data assimilation: Operational, sequential and variational*, J. Meteor. Soc. Japan **75** (1997), 181–189.

[110] A. Jazwinski, *Stochastic processes and filtering theory*, Cambridge, MA: Academic Press, 1970.

[111] C. Johnson, N. K. Nichols, and B. J. Hoskins, *Very large inverse problems in atmosphere and ocean modelling*, Internat. J. Numer. Methods Fluids **47** (2005), 759–771.

[112] K. Judd, *Forecasting with imperfect models, dynamically constrained inverse problems, and gradient descent algorithms*, Physica D: Nonlinear Phenomena **237** (2008), 216–232.

[113] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, Vol **160**, Springer Science & Business Media, Berlin, 2006.

[114] R. E. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic Eng. **82** (1960), 35–45.

[115] E. J. Kostelich et al., *Accurate state estimation from uncertain data and models: An application of data assimilation to mathematical models of human brain tumors*, Biol. Direct **6** (2011), 64.

[116] A. Kovalenko, T. Mannseth, and G. Nævdal, *Error estimate for the ensemble Kalman filter analysis step*, SIAM J. Matrix Anal. Appl. **32** (2011), 1275–1287.

[117] P. Lancaster and L. Rodman, *Algebraic Riccati equations*, Oxford: Clarendon Press, 1995.

[118] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, United States Governm. Press Office, Los Angeles, CA, 1950.

[119] K. Law, A. M. Stuart, and K. Zygalakis, *Data assimilation*, in *Texts in Applied Mathematics. A Mathematical Introduction*, Vol **62**, Springer, Cham, 2015.

[120] A. S. Lawless, *Variational data assimilation for very large environmental problems*, in *Large scale inverse problems, Radon Ser. Comput. Appl. Math.*, Vol **13**, De Gruyter, Berlin, 2013, 55–90.

[121] A. S. Lawless, S. Gratton, and N. K. Nichols, *Approximate iterative methods for variational data assimilation*, Internat. J. Numer. Methods Fluids **47** (2005), 1129–1135.

[122] A. S. Lawless, S. Gratton, and N. K. Nichols, *An investigation of incremental 4d-var using non-tangent linear models*, Q. J. R. Meteorol. Soc. **131** (2005), 459–476.

[123] A. S. Lawless and N. K. Nichols, *Inner-loop stopping criteria for incremental four-dimensional variational data assimilation*, Mon. Weather Rev. **134** (2006), 3425–3435.

[124] A. S. Lawless et al., *Approximate Gauss-Newton methods for optimal state estimation using reduced-order models*, Internat. J. Numer. Methods Fluids **56** (2008), 1367–1373.

[125] A. S. Lawless et al., *Using model reduction methods within incremental four-dimensional variational data assimilation*, Mon. Weather Rev. **136** (2008), 1511–1522.

[126] L. M. Lawson et al., *A data assimilation technique applied to a predator-prey model*, Bull. Math. Biol. **57** (1995), 593–617.

[127] Z. Li and I. M. Navon, *Optimality of variational data assimilation and its relationship with the Kalman filter and smoother*, Q. J. R. Meteorol. Soc. **127** (2001), 661–683.

[128] C. Lieberman et al., *Hessian-based model reduction: Large-scale inversion and prediction*, Internat. J. Numer. Methods Fluids **71** (2013), 135–150.

[129] D. M. Livings, S. L. Dance, and N. K. Nichols, *Unbiased ensemble square root filters*, Physica D: Nonlinear Phenomena **237** (2008), 1021–1028.

[130] Y. Luo et al., *Ecological forecasting and data assimilation in a data-rich era*, Ecol. Appl. **21** (2011), 1429–1442.

[131] J. Mandel et al., *Hybrid Levenberg-Marquardt and weak-constraint ensemble Kalman smoother method*, Nonlinear Proc. Geoph. **23** (2016), 59–73.

[132] B. C. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control **26** (1981), 17–32.

[133] J. L. Morales and J. Nocedal, *Automatic preconditioning by limited memory quasi-newton updating*, SIAM J. Optim. **10** (2000), 1079–1096.

[134] J. L. Mueller and S. Siltanen, *Linear and nonlinear inverse problems with practical applications*, in *Computational science and engineering*, Vol **10**, SIAM, New York, 2012.

[135] I. M. Navon, *Data assimilation for numerical weather prediction: A review*, in *Data assimilation for atmospheric, oceanic and hydrologic applications*, Springer, Berlin, 2009, 21–65.

[136] L. Nerger and W. Hiller, *Software for ensemble-based data assimilation systems—Implementation strategies and scalability*, Computers & Geosciences, **55** (2013), 110–118.

[137] L. Nerger, W. Hiller, and J. Schröter, *A comparison of error subspace Kalman filters*, Tellus A **57** (2005), 715–735.

[138] A. Neumaier, *Solving ill-conditioned and singular linear systems: A tutorial on regularization*, SIAM Rev. **40** (1998), 636–666.

[139] N. K. Nichols, *Data assimilation: Aims and basic concepts*, in *Data Assimilation for the Earth System*, Springer, Berlin, 2003, 9–20.

[140] S. Niu et al., *The role of data assimilation in predictive ecology*, Ecosphere **5** (2014), 1–16.

[141] J. Nocedal and S. Wright, *Numerical optimization*, Springer Science & Business Media, Berlin, 2006.

[142] E. Ott et al., *A local ensemble Kalman filter for atmospheric data assimilation*, Tellus A **56** (2004), 415–428.

[143] S. K. Park and L. Xu, *Data assimilation for atmospheric, oceanic and hydrologic applications*, Vol **2**, Springer Science & Business Media, Berlin, 2013.

[144] B. Peherstorfer and K. Willcox, *Dynamic data-driven reduced-order models*, Comput. Methods Appl. Mech. Engrg. **291** (2015), 21–41.

[145] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, SIAM Rev. **60** (2018), 550–591.

[146] D. T. Pham, J. Verron, and M. C. Roubaud, *A singular evolutive extended Kalman filter for data assimilation in oceanography*, J. Marine Syst. **16** (1998), 323–340.

[147] V. Rao and A. Sandu, *A time-parallel approach to strong-constraint four-dimensional variational data assimilation*, J. Comput. Phys. **313** (2016), 583–593.

[148] V. Rao et al., *Robust data assimilation using l_1 and Huber norms*, SIAM J. Sci. Comput. **39** (2017), B548–B570.

[149] S. Reich and C. Cotter, *Probabilistic forecasting and Bayesian data assimilation*, Cambridge University Press, Cambridge, MA, 2015.

[150] S. Reich and C. J. Cotter, *Ensemble filter techniques for intermittent data assimilation*, in *Large scale inverse problems, Radon Ser. Comput. Appl. Math.*, Vol **13**, De Gruyter, Berlin, 2013, 91–134.

[151] C. J. Rhodes and T. D. Hollingsworth, *Variational data assimilation with epidemic models*, J. Theor. Biol. **258** (2009), 591–602.

[152] C. Robert et al., *A reduced-order strategy for 4D-Var data assimilation*, J. Marine Syst. **57** (2005), 70–82.

[153] D. Rozier et al., *A reduced-order Kalman filter for data assimilation in physical oceanography*, SIAM Rev. **49** (2007), 449–465.

[154] A. K. Saibaba, J. Chung, and K. Petroske, Quantifying uncertainties in large-scale Bayesian linear inverse problems using Krylov subspace methods, arXiv preprint arXiv:1808.09066, (2018).

[155] Y. Sasaki, *An objective analysis based on the variational method*, J. Meteor. Soc. Japan **36** (1958), 77–88.

[156] Y. Sasaki, *Some basic formalisms in numerical variational analysis*, Mon. Weather Rev. **98** (1970), 875–883.

[157] S. J. Schiff, *Kalman meets neuron: The emerging intersection of control theory with neuroscience*, in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN: IEEE, 2009, 3318–3321.

[158] T. Schuster, B. Hahn, and M. Burger, *Dynamic inverse problems: Modelling-regularization-numerics*, Inverse Problems **34** (2018), 040301.

[159] B. Schwartz, S. Gannot, and E. A. P. Habets, *Online speech dereverberation using Kalman filter and EM algorithm*, IEEE/ACM Trans. Audio, Speech, and Language Process. **23** (2014), 394–406.

[160] A. Spantini et al., *Optimal low-rank approximations of Bayesian linear inverse problems*, SIAM J. Sci. Comput. **37** (2015), A2451–A2487.

[161] R. Ştefănescu, A. Sandu, and I. M. Navon, *POD/DEIM reduced-order strategies for efficient four dimensional variational data assimilation*, J. Comput. Phys. **295** (2015), 569–595.

[162] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer. **19** (2010), 451–559.

[163] J. M. Tabeart et al., *The conditioning of least-squares problems in variational data assimilation*, Numer. Linear Algebra Appl. **25** (2018), e2165.

[164] O. Talagrand and P. Courtier, *Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory*, Q. J. R. Meteorol. Soc. **113** (1987), 1311–1328.

[165] M. K. Tippett et al., *Ensemble square root filters*, Mon. Weather Rev. **131** (2003), 1485–1490.

[166] Y. Trémolet, *Diagnostics of linear and incremental approximations in 4D-Var*, Q. J. R. Meteorol. Soc. **130** (2004), 2233–2251.

[167] Y. Trémolet, *Incremental 4D-Var convergence study*, Tellus A **59** (2007), 706–718.
[168] Y. Trémolet, *Model-error estimation in 4D-Var*, Q. J. R. Meteorol. Soc. **133** (2007), 1267–1280.
[169] Y. Trémolet, *Computation of observation sensitivity and observation impact in incremental variational data assimilation*, Tellus A **60** (2008), 964–978.
[170] J. Tshimanga et al., *Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation*, Q. J. R. Meteorol. Soc. **134** (2008), 751–769.
[171] P. J. van Leeuwen, *Nonlinear data assimilation in geosciences: An extremely efficient particle filter*, Q. J. R. Meteorol. Soc. **136** (2010), 1991–1999.
[172] M. Verlaan and A. W. Heemink, *Tidal flow forecasting using reduced rank square root filters*, Stoch. Hydrol. Hydraul. **11** (1997), 349–368.
[173] P. T. M. Vermeulen and A. W. Heemink, *Model-reduced variational data assimilation*, Mon. Weather Rev. **134** (2006), 2888–2899.
[174] J. Verron et al., *An extended Kalman filter to assimilate satellite altimeter data into a nonlinear numerical model of the tropical Pacific Ocean: Method and validation*, J. Geophys. Res. **104** (1999), 5441–5458.
[175] C. R. Vogel, *Computational methods for inverse problems*, Vol **23**, Philadelphia: SIAM, 2002.
[176] C. K. Wikle and L. M. Berliner, *A Bayesian tutorial for data assimilation*, Phys. D **230** (2007), 1–16.

How to cite this article: Freitag MA. Numerical linear algebra in data assimilation. *GAMM-Mitteilungen*. 2020;43:e202000014. https://doi.org/10.1002/gamm.202000014