Received: 05 October 2017 | Revised: 23 May 2018 | Accepted: 30 May 2018 | Published on: 12 October 2018

DOI: 10.1002/qj.3347

RESEARCH ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

Information constraints in variational data assimilation

Michael Kahnert[1,2]

[1]Research Department, Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 601 76 Norrköping, Sweden
[2]Department of Space, Earth and Environment, Chalmers University of Technology, Maskingränd 2, 412 96 Gothenburg, Sweden

Correspondence
Michael Kahnert, Research Department, Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 601 76 Norrköping, Sweden.
Email: michael.kahnert@smhi.se

Data assimilation of indirect observations from remote-sensing instruments often leads to highly under-determined inverse problems. Here a formulation of the variational method is discussed in which (a) the information content of the observations is systematically analysed by methods borrowed from retrieval theory; (b) the model space is transformed into a phase space in which one can partition the model variables into those that are related to the degrees of freedom for signal and noise, respectively; and (c) the minimization routine in the variational analysis is constrained to act on the signal-related phase-space variables only. This is done by truncating the dimension of the phase space. A first test of the method indicates that the constrained analysis speeds up computation time by about an order of magnitude compared with the formulation without information constraints.

KEYWORDS
chemical transport modelling, data assimilation, inverse problems, remote sensing

1 | INTRODUCTION

Data assimilation is an essential part of many environmental forecasting and analysis systems. The basic mathematical concepts of data assimilation are well understood and described in various textbooks and reviews (e.g. Swinbank *et al.*, 2003; Lahoz *et al.*, 2010). Applications include meteorological (Rabier, 2005; Bannister, 2017), oceanographic (Evensen, 2007), and chemical-transport models (Sandu and Chai, 2011).

One of the specific applications in data assimilation is related to the exploitation of indirect observations. A typical example is the assimilation of aerosol optical depth (AOD; e.g. Schutgens *et al.*, 2010; Liu *et al.*, 2011; Saide *et al.*, 2013; 2014; Chen *et al.*, 2014; Pagowski *et al.*, 2014; Rubin and Collins, 2014) or aerosol backscattering measurements (e.g. Wang *et al.* 2013; 2014a; 2014b; Pagowski *et al.*, 2014; Zhang *et al.* 2011, 2014) from remote-sensing instruments in an aerosol transport model. A characteristic of such applications is that the number of observed parameters is often significantly smaller than the number of model variables. Several approaches have been considered in which the number of control variables in the assimilation has been reduced accordingly.

Liu *et al.* (2011) assimilated AOD by allowing the analysis algorithm to independently adjust all aerosol components in all size bins and model layers. Thus, this approach makes no attempt to reduce the dimension of the control space in accordance with the small number of signal degrees of freedom. Benedetti *et al.* (2009) performed AOD assimilation by using a single control variable per grid point (the total aerosol mass mixing ratio). Wang *et al.* (2014a; 2014b) assimilated lidar observations, using either PM_{10} or both $PM_{2.5}$ and $PM_{2.5-10}$ as control variables.[1] Saide *et al.* (2013) used a limited number of control variables, namely the total aerosol mass mixing ratio per size bin.

The idea of reducing the dimension of the control space according to the number of degrees of freedom for signal can be generalized with standard methods borrowed from retrieval theory (Rodgers, 2000). In the example of aerosol optical observations, a systematic approach takes into account the physical relations between the aerosols' physical, chemical, and morphological properties and their optical properties, which are encoded in the aerosol optics observation operator. Based on this relation, one can analyse the information content of the observations and determine which model variables – or linear combination thereof – are most strongly

[1]PM_{10} denotes particulate matter with a diameter up to 10 μm. Similarly, $PM_{2.5}$ refers to fine particulate matter with a diameter up to 2.5 μm. Coarse particulate matter with a diameter between 2.5 and 10 μm is denoted by $PM_{2.5-10}$.

related to the information content of the measurements (e.g. Kahnert and Andersson, 2017). The minimization in the variational analysis can then be constrained to act on those model variables only.

Analysis of the information content of observations can be a useful tool in remote sensing retrieval methods (Rodgers, 2000). For instance, the information content of lidar observations has been investigated with regard to different retrieval products (Veselovskii *et al.*, 2004; 2005; Burton *et al.*, 2016).

One way of analyzing the information content of observations is to normalize the Jacobian of the observation operator with the error covariance matrices of the background and observations, and to perform a singular value decomposition (SVD) of the normalized Jacobian. The singular values allow one to compute the number of degrees of freedom for signal. This provides us with an estimate for the number of model variables that can be constrained by the measurements. The SVD approach has been applied in different applications (Joiner and da Silva, 1998; Rabier *et al.*, 2002; Johnson *et al.*, 2005a; 2005b; Bocquet, 2009; 2011). For instance, a quantification of information contents has been used in numerical weather forecasting to reduce the effect of prior information in the analysis (Joiner and da Silva, 1998). In Rabier *et al.* (2002) the SVD approach has been used to select those channels from infrared sounding instruments with the highest information content. The SVD method has also been applied to obtain a deeper understanding of filtering and interpolation aspects of four-dimensional variational data analysis (4D-Var) assimilation systems (Johnson *et al.*, 2005a; 2005b). Bocquet (2009) optimized the spatial resolution in geophysical models, which resulted in a Bayesian theory for optimal discretization of control space (Bocquet *et al.*, 2011). In Bocquet and Wu (2011) analytical solutions of this optimization problem in the asymptotic limit of large grids have been derived. This derivation also made use of the SVD of the normalized Jacobian. Kahnert and Andersson (2017) performed an SVD analysis of the information content of lidar observations and incorporated this into a variational analysis algorithm in order to constrain the minimization of the cost function to act on the signal-related model variables only.

Other methods of analysing information content of observations have been applied. For instance, in the Bayesian theory of optimal discretization of control space, Bocquet *et al.* (2011) used three different information criteria – the Fisher criterion, the degrees of freedom for signal, and the relative entropy gain – to test the optimality of the choice of grid. Cardinali *et al.* (2004) computed the influence matrix to compute diagnostics of the impact of observations. Xu (2006) compared two metrics of information gain, the relative entropy and the Shannon entropy difference, to quantify the information content of radar data in a coupled atmosphere–ocean model.

In the work presented here, an SVD analysis will be tested to exploit information constraints in a 3D-Var algorithm. By use of the SVD the control space is partitioned into model variables (or, rather, linear combinations thereof) that can be related to the degrees of freedom for signal, and those that can be related to the degrees of freedom for noise. The minimization of the cost function is constrained to act on the former components only. The approach taken here will overcome the *ad hoc* elements and approximations made in Kahnert and Andersson (2017). In the latter, the SVD was done in physical space. In that case the leading dimension of the normalized Jacobian of the observation operator is very high, which makes the SVD numerically infeasible. For this reason, Kahnert and Andersson (2017) performed an *ad hoc* reduction of physical space. They performed the SVD only at the observation points, and then applied it throughout the model domain. Such an approximation will no longer be necessary here, as the SVD will be performed in spectral space, which has a much lower dimension than physical space. Further, in Kahnert and Andersson (2017) the constraints were introduced into the minimization of the cost function by adding an extra term to the cost function. Here, the constraints will be implemented by truncating the dimension of the control space in such a way that only the signal-related variables are retained. By contrast to the approach in Kahnert and Andersson (2017), this leads to a fully Bayesian formulation of the cost function. It will be seen that the truncation of the control space to the signal-related variables is a rather mild approximation that gives an analysis very close to the unconstrained analysis, but at much reduced numerical costs; the resulting formulation of variational data analysis reduces CPU time requirements by about one order of magnitude. No such reduction in computation time was obtained in the approach in Kahnert and Andersson (2017).

The method will be illustrated by performing an observing system simulation experiment (OSSE) for lidar observations. Two different aerosol optics models will be employed as observation operators. First, a simple linear model will be used, which has been tested earlier (Kahnert, 2008; Kahnert and Andersson, 2017). Second, a recently developed nonlinear aerosol optics model will be applied Andersson and Kahnert (2016). That model accounts not only for homogeneous internal mixing, but also for the non-spherical fractal aggregate morphology of black carbon, and for the inhomogeneous internal structure of black carbon mixed with soluble compounds. The latter is described by the recently developed core grey-shell model (Kahnert *et al.*, 2013), which is significantly more versatile than the conventional core-shell model (e.g. Jacobson, 2000). This nonlinear optics model has, so far, not yet been tested as part of a data analysis algorithm.

The target audience of this paper are model developers and those with an interest in data assimilation application. The theoretical background is given in section 2. The OSSE is presented in section 3. A discussion of possible applications to 4D-Var and the ensemble Kalman filter can be found in section 4. Concluding remarks are given in section 5.

2 | THEORY

2.1 | Notation

A CTM typically considers a certain model domain (such as Europe, the Northern Hemisphere, or the entire globe), which is discretized into a three-dimensional grid. Let us assume that the grid consists of N_x horizontal (longitudinal) and N_y vertical (latitudinal) grid lines, as well as N_z vertical layers. The chemical state of the atmosphere is characterized by the mixing ratios of N_c trace gas and aerosol components (where the latter are typically sorted into different size bins), each of which are specified in each grid point. Thus the total dimension of our state space is $N = N_x N_y N_z N_c$. Suppose we collect all mixing ratios from all grid points into a vector $\mathbf{x}$. Let us further suppose that we have a set of N_{obs} observations that we collect into a vector $\mathbf{y}$. Finally, let us suppose that we have a forward model represented by the observation operator

$$\mathcal{H} : (\overline{\mathbb{R}}_+)^N \to \mathbb{R}^{N_{obs}}, \mathbf{x} \mapsto \mathcal{H}(\mathbf{x}), \tag{1}$$

which maps from model to observation space.[2] Ideally, the relation between the observations and the state vector would be given by $\mathbf{y} = \mathcal{H}(\mathbf{x})$. However, in a realistic treatment of inverse problems, it is essential to account for the fact that there are several sources of error and uncertainty, which limit the accuracy with which we can relate the model state to the measurements. Thus we have

$$\mathbf{y} = \mathcal{H}(\mathbf{x}) + \boldsymbol{\epsilon}_o, \tag{2}$$

where $\boldsymbol{\epsilon}_o$ denotes the vector of observation errors. The latter encompasses all sources of error entering into this relation, such as the measurement error, representativeness errors (Janjić *et al.*, 2017), and errors introduced by approximations and assumptions in the observation operator.

It will henceforth be assumed that a first-order expansion of the observation operator is sufficiently accurate for our purposes. Thus, a Taylor expansion in the neighbourhood of a point $\mathbf{x}_b$ gives:

$$\mathcal{H}(\mathbf{x}) = \mathcal{H}(\mathbf{x}_b) + \mathbf{H}\delta\mathbf{x}, \tag{3}$$

where $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}_b$, and where $\mathbf{H}$ denotes the Jacobian of the operator $\mathcal{H}$.

In the following sections it will sometimes be useful to explicitly write vectors and matrices in terms of their components. Vector components will be denoted by contra-variant tensor components with upper indices, such as

$$\begin{aligned} x^{i,j,l,k} \quad & i = 1, \dots, N_x, j = 1, \dots, N_y, l = 1, \dots, N_z, \\ & k = 1, \dots, N_c; \\ y^p \quad & p = 1, \dots, N_{obs}. \end{aligned} \tag{4}$$

The transpose $\mathbf{x}^T$ of a vector is a dual- or co-vector, represented by covariant tensor components with lower indices, given in components by $x_{i,j,l,k}$. Matrix–vector products are written in the form

$$(\mathbf{H}\delta\mathbf{x})^p = H^p{}_{i,j,l,k} x^{i,j,l,k}, \tag{5}$$

where a summation over repeated co- and contra-variant indices is implied (unless otherwise stated).

In the following we will need error variances and covariances for the observation errors $\boldsymbol{\epsilon}_o$, which are expressed by the observation-error covariance matrix

$$\mathbf{R} = \overline{\boldsymbol{\epsilon}_o \boldsymbol{\epsilon}_o^T}, \tag{6}$$

where the overbar denotes the arithmetic mean over an ensemble of observation errors. The expression over which the ensemble mean is evaluated is a dyadic product, i.e. the observation-error covariance matrix has components

$$R^p{}_{p'} = \overline{(\epsilon_o)^p (\epsilon_o)_{p'}}, \qquad p, p' = 1, \dots, N_{obs}. \tag{7}$$

By definition, the error covariance matrix is symmetric and positive definite.

Similarly, we will need a background estimate (or *a priori*) $\mathbf{x}_b$ of the chemical state of the atmosphere, which is usually given by a model forecast, and we need the error variances and covariance of the corresponding background errors $\boldsymbol{\epsilon}_b$, which are given by the background-error covariance matrix

$$\mathbf{B} = \overline{\boldsymbol{\epsilon}_b \boldsymbol{\epsilon}_b^T}, \tag{8}$$

or, in components,

$$B^{i,j,l,k}{}_{i',j',l',k'} = \overline{(\epsilon_b)^{i,j,l,k} (\epsilon_b)_{i',j',l',k'}}. \tag{9}$$

2.2 | Variational data analysis

We consider the conditional probability distribution function that the atmosphere is in state $\mathbf{x}$ given observations $\mathbf{y}$, denoted by $P(\mathbf{x} \mid \mathbf{y})$. According to Bayes' theorem,

$$P(\mathbf{x} \mid \mathbf{y}) \propto P(\mathbf{x})\, P(\mathbf{y} \mid \mathbf{x}), \tag{10}$$

where $P(\mathbf{x})$ is the prior probability distribution function (PDF) of the state $\mathbf{x}$, and $P(\mathbf{y} \mid \mathbf{x})$ is the conditional PDF of $\mathbf{y}$ given $\mathbf{x}$. Assuming Gaussian statistics, we have

$$P(\mathbf{x}) = \frac{1}{\mid 2\pi\mathbf{B} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b)\right] \tag{11}$$

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{\mid 2\pi\mathbf{R} \mid^{1/2}} \exp\left[-\frac{1}{2}\{\mathcal{H}(\mathbf{x}) - \mathbf{y}\}^T \mathbf{R}^{-1} \{\mathcal{H}(\mathbf{x}) - \mathbf{y}\}\right]. \tag{12}$$

Equations 10–12 can be summarized as

$$P(\mathbf{x} \mid \mathbf{y}) \propto \exp[-J(\mathbf{x})], \tag{13}$$

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + \frac{1}{2}\{\mathcal{H}(\mathbf{x}) - \mathbf{y}\}^T \mathbf{R}^{-1} \{\mathcal{H}(\mathbf{x}) - \mathbf{y}\}, \tag{14}$$

where J is the cost function. (In probability theory, J is also known as the negative log-likelihood.) The data analysis problem consists of finding that state $\mathbf{x}_a$ for which $P(\mathbf{x} \mid \mathbf{y})$ has its maximum. Thus the *analysis* or *analysed state* $\mathbf{x}_a$

[2] Note that the mixing ratios in model space cannot have negative values. Therefore, the model states are elements of $(\overline{\mathbb{R}}_+)^N$, where $\mathbb{R}_+$ denotes the set of positive real numbers, and its closure is given by $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{0\}$.

represents the most probable state of the system, all available information and error statistics considered. In the variational method, the maximum of $P(\mathbf{x} \mid \mathbf{y})$ is determined by finding the minimum of J. This is achieved iteratively by a descent algorithm that makes use of the gradient ∇J. The term *data analysis* refers to the solution $\mathbf{x}_a$ at a specific time. The term *data assimilation* refers to the process in which one merges information from observations and a dynamic model. For instance, in 3D-Var assimilation, this is done sequentially; whenever observations are available, the most recent model forecast is being updated by computing the analysis $\mathbf{x}_a$. The latter is used as the initial state for producing the next forecast.

By Taylor expansion of the observation operator according to Equation 3, one arrives at the incremental formulation of variational data analysis

$$J(\delta\mathbf{x}) = \frac{1}{2}\delta\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x} + \frac{1}{2}(\mathbf{H}\delta\mathbf{x} - \delta\mathbf{y})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{x} - \delta\mathbf{y}), \quad (15)$$

where $\delta\mathbf{y} = \mathbf{y} - \mathcal{H}(\mathbf{x}_b)$. The control variable $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}_b$ is iteratively varied until the minimum of the cost function is found. The analysis increment $\delta\mathbf{x}_a$ is the amount by which the analysis updates the *a priori* model estimate $\mathbf{x}_b$, i.e. $\mathbf{x}_a = \mathbf{x}_b + \delta\mathbf{x}_a$.

The previous expression can be further streamlined. The matrix $\mathbf{B}^{-1}$ can be decomposed as $\mathbf{B}^{-1} = \mathbf{A}^{\mathrm{T}} \cdot \mathbf{A}$. The choice of $\mathbf{A}$ is not unique. For instance, one can take $\mathbf{A} = \mathbf{B}^{-1/2}$, where $\mathbf{B}^{1/2}$ denotes the positive square root of $\mathbf{B}$. (For a positive definite symmetric matrix, the positive square root exists and is unique.) The data analysis code that will be applied in section 3 is a spectral code, in which a decomposition into spectral eigenvalues is being applied to the horizontal dimensions of control space. Let us denote the spectral decomposition by $\mathbf{B}^{-1} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$ (e.g. Kahnert, 2008), and let us use this decomposition in the following (although any other decomposition would work equally well). Similarly, we decompose $\mathbf{R}^{-1} = \mathbf{R}^{-\mathrm{T}/2}\mathbf{R}^{-1/2}$, where $\mathbf{R}^{-\mathrm{T}/2}$ denotes the transpose of $\mathbf{R}^{-1/2}$. By

$$\delta\widetilde{\mathbf{x}} = \mathbf{U}\delta\mathbf{x}, \quad (16)$$

$$\delta\widetilde{\mathbf{y}} = \mathbf{R}^{-1/2}\delta\mathbf{y}, \quad (17)$$

$$\widetilde{\mathbf{H}} = \mathbf{R}^{-1/2}\mathbf{H}\mathbf{U}^{-1}, \quad (18)$$

we obtain

$$J(\delta\widetilde{\mathbf{x}}) = \frac{1}{2}\delta\widetilde{\mathbf{x}}^{\mathrm{T}}\delta\widetilde{\mathbf{x}} + \frac{1}{2}(\widetilde{\mathbf{H}}\delta\widetilde{\mathbf{x}} - \delta\widetilde{\mathbf{y}})^{\mathrm{T}}(\widetilde{\mathbf{H}}\delta\widetilde{\mathbf{x}} - \delta\widetilde{\mathbf{y}}). \quad (19)$$

The gradient of the cost function, which is needed in the iterative minimization algorithm, is given by

$$\nabla J(\delta\widetilde{\mathbf{x}}) = \delta\widetilde{\mathbf{x}} + \widetilde{\mathbf{H}}^{\mathrm{T}}(\widetilde{\mathbf{H}}\delta\widetilde{\mathbf{x}} - \delta\widetilde{\mathbf{y}}). \quad (20)$$

In the minimization algorithm, $\delta\widetilde{\mathbf{x}}$ is iteratively optimized. Thus the matrix-vector products in Equations 19 and 20 have to be computed in each step of the iteration, which can become very time-consuming. The approach reviewed in the following section results in a diagonalization of the observation operator as well as a reduction of the dimension of the control space. The goal is to test whether or not this approach can help us to reduce the computation time in the variational method.

2.3 | Information constraints

We want to recast variational data analysis into a form in which the minimization algorithm adjusts only those model variables for which the observations carry information. This can be achieved by performing a singular value decomposition (SVD) of the scaled observation operator in Equation 18, i.e.

$$\widetilde{\mathbf{H}} = \mathbf{V}_{\mathrm{L}}\mathbf{H}'\mathbf{V}_{\mathrm{R}}^{\mathrm{T}}, \quad (21)$$

where $\mathbf{V}_{\mathrm{L}}$ and $\mathbf{V}_{\mathrm{R}}$ are orthogonal matrices containing the left- and right-singular vectors, respectively, and $\mathbf{H}'$ is a diagonal matrix containing the singular values. Those singular values that are larger than about unity correspond to degrees of freedom for signal, i.e. to model variables for which the observations carry a significant amount of information. The SVD can be employed to transform model space from the concrete physical space of aerosol mixing ratios into a more abstract phase space, in which the state-vector components can be partitioned into signal- and noise-related components. The transformation is given by

$$\delta\mathbf{x}' = \mathbf{V}_{\mathrm{R}}^{\mathrm{T}}\delta\widetilde{\mathbf{x}} = \mathbf{V}_{\mathrm{R}}^{\mathrm{T}}\mathbf{U}\delta\mathbf{x}. \quad (22)$$

Those components of $\delta\mathbf{x}'$ that belong to singular values on the order of unity or larger can be controlled by the measurements. All other components of $\delta\mathbf{x}'$ can be left unchanged by the minimization algorithm, as the measurements carry insufficient information on those components.

For more information on the underlying theory, the reader is referred to Rodgers (2000) (chapter 2). However, one can obtain an intuitive understanding of the general idea by inspecting Equation 18. If we want to know for which model variables the observations contain information, then we have to investigate the observation operator, because this operator encodes the relation between modelled and measured variables. It is also reasonable that this operator is scaled by $\mathbf{R}^{-1/2}$ and $\mathbf{U}^{-1}$ (or $\mathbf{B}^{1/2}$), because this accounts for the fact that the information contents of the observations is larger when the observation-error standard deviations are smaller in relation to those of the *a priori* estimate. Also, the scaling makes the elements of this matrix dimensionless. In the simple case of a single model variable and a single direct observation, the expression in Equation 18 would simply become $\widetilde{H} = \sigma_b/\sigma_o$, where σ_b and σ_o are the error standard deviations of the background estimate and of the observation, respectively. An observation contains a significant amount of information that can improve the model estimate if $\sigma_b/\sigma_o > 1$. In a more general multidimensional problem with indirect observations, the singular values of the matrix $\widetilde{\mathbf{H}}$ become the generalization of the ratio σ_b/σ_o (e.g. Kahnert and Andersson, 2017).

We now perform the following change of variables in the analysis algorithm:

$$\delta\mathbf{x}' = \mathbf{V}_{\mathrm{R}}^{\mathrm{T}}\delta\widetilde{\mathbf{x}}, \quad (23)$$

$$\delta\mathbf{y}' = \mathbf{V}_{\mathrm{L}}^{\mathrm{T}}\delta\widetilde{\mathbf{y}}, \quad (24)$$

$$\mathbf{H}' = \mathbf{V}_{\mathrm{L}}^{\mathrm{T}}\widetilde{\mathbf{H}}\mathbf{V}_{\mathrm{R}}. \quad (25)$$

Substitution into Equations 19 and 20 yields

$$J(\delta\mathbf{x}') = \frac{1}{2}\delta\mathbf{x}'^{\mathrm{T}}\delta\mathbf{x}' + \frac{1}{2}(\mathbf{H}'\delta\mathbf{x}' - \delta\mathbf{y}')^{\mathrm{T}}(\mathbf{H}'\delta\mathbf{x}' - \delta\mathbf{y}'), \quad (26)$$

$$\nabla J = \delta\mathbf{x}' + \mathbf{H}'^{\mathrm{T}}(\mathbf{H}'\delta\mathbf{x}' - \delta\mathbf{y}'). \quad (27)$$

Once the analysis increment $\delta\mathbf{x}'_{\mathrm{a}}$ has been found, the transformation from phase space to physical space is performed by inverting Equations 23 and 16, i.e.

$$\delta\mathbf{x}_{\mathrm{a}} = \mathbf{U}^{-1}\mathbf{V}_{\mathrm{R}}\delta\mathbf{x}'_{\mathrm{a}}. \quad (28)$$

The formulation in Equations 26 and 27 has potential advantages over that in Equations 19 and 20:

- The new control vector $\delta\mathbf{x}'$ is an element of a space in which the components can be partitioned into signal- and noise-related components. This allows us to constrain the minimization algorithm to act on the former vector components only. To be more specific, suppose that we have N_0 non-zero singular values of the matrix $\mathbf{H}'$ that are arranged so that $h'_1 > h'_2 > \cdots > h'_k > 1$ and $1 > h'_{k+1} > \cdots > h'_{N_0}$, where $N_0 \leq \min\{N, N_{\mathrm{obs}}\}$, and where equality holds if $\mathrm{rank}\mathbf{H}' = \min\{N, N_{\mathrm{obs}}\}$. Then we truncate the phase space to the elements $\delta x'^1, \cdots, \delta x'^k$. These elements can be freely adjusted in the minimization of the cost function, while all other elements are fixed, $\delta x'^i = 0$, $i = k+1, \ldots, N$. In other words, we constrain the algorithm to act on the signal-related phase space variables only. For brevity, let us refer to these as *information constraints*. This truncation will reduce CPU time requirements, since it reduces the dimension of the space in which the minimization is performed. In fact, even if we truncate the control vector at $N_{\mathrm{cut}} = N_0$, then we would still substantially reduce the dimension of control space, because typically $N_0 \ll N$.
- Equations 19 and 20 involve time-consuming matrix-vector multiplications with the matrix $\widetilde{\mathbf{H}}$, while Equations 26 and 27 only involve multiplications with the *diagonal* matrix $\mathbf{H}'$. These multiplications have to be performed in each iteration step of the minimization algorithm. Typically, the algorithm may loop through hundreds of iterations before convergence is reached. Therefore, we can expect the algorithm based on Equations 26 and 27 to be significantly faster. The only investment that is required is the SVD in Equation 21. However, since the SVD is computed outside the iterative minimization procedure, this can be expected to be a minor investment compared with the prospective gain in CPU time.

3 | TESTING OF THE ALGORITHM

3.1 | MATCH 3D-Var modelling system

The constrained analysis algorithm will be illustrated by use of a chemical transport model with a spectral data analysis code. The Multiple scale Atmospheric Transport and CHemistry modelling system (MATCH; Andersson *et al.*, 2007; 2015) with its spectral 3D-Var module (Kahnert, 2008) will be used. The model contains a photochemistry mass-transport model with 64 species. The tests are performed with two different model versions for aerosols.

1. The simplest version contains four secondary inorganic aerosol (SIA) species (ammonium sulphate, ammonium nitrate, other sulphates, and other nitrates) and 16 primary aerosol components, namely, mineral dust, sea salt, organic carbon, and elemental carbon, each in four size bins, and each assumed not to participate in any chemical reactions. Aerosol microphysical processes are switched off. The four size bins comprise the particle-radius intervals [10, 50], [50, 500], [500, 1250], and [1250, 5000] nm. Thus the 20 control variables (in each grid cell) are:

1–4	Organic carbon (OC) in size bins 1–4
5–8	Black carbon (BC) in size bins 1–4
9–12	Dust in size bins 1–4
13–16	Sea salt in size bins 1–4
17	Ammonium sulphate
18	Ammonium nitrate
19	Other sulphates
20	Other nitrates

   A simple aerosol optics model based on assuming all aerosol species to be externally mixed homogeneous spheres is implemented in this version (Kahnert, 2008). The primary aerosol particles are emitted in four different size bins and remain in their respective bins while undergoing transport and deposition. The four SIA components are described by their total mass concentration. In the optics model, the total SIA mass is distributed among the four size bins in the proportions 0.1 : 0.6 : 0.2 : 0.1. Owing to the external-mixture approximation, the observation operator is linear.
2. The second version (Andersson *et al.*, 2015) uses the same photochemistry scheme as the first one, but with microphysical processes switched on. The latter are computed in the aerosol microphysics model Sectional Aerosol module for Large-Scale Applications (SALSA; Kokkola *et al.*, 2008). The model contains a total of 20 size bins, each representing different size ranges and aerosol mixing states. The total size range covers particle radii from 1.5,nm to 5 μm. In total, the aerosol state is described by 76 variables per grid cell, representing different chemical aerosol components in different size bins. Table 1 (taken from Andersson and Kahnert (2016)) gives an overview of all 76 control variables.

   In this version the optics model accounts for internal mixing of different aerosol species; this results in a nonlinear observation operator. Its Jacobian is computed by numerical differentiation at the expansion point. Also, for BC particles mixed internally with liquid-phase compounds, such as OC and secondary inorganic species, the

TABLE 1 Size bins and chemical species in the MATCH-SALSA aerosol microphysical transport model. An "x" indicates that the species is present in that particular size bin (based on Andersson and Kahnert, 2016)

Size bin	r (nm)	Mixing state	OC	BC	Dust	Sea salt	PSO_x	PNO_x	PNH_x
1	1.5–3.8	internal	x				x		x
2	3.8–9.8	internal	x				x		x
3	9.8–25	internal	x				x		x
4	25–49	internal+H_2O	x	x	x	x	x		x
5	49–96	internal+H_2O	x	x	x	x	x		x
6	96–187	internal+H_2O	x	x	x	x	x		x
7	187–350	internal+H_2O	x	x	x	x	x		x
8	25–49	external	x	x			x		x
9	49–96	external	x	x			x		x
10	96–187	external	x	x			x		x
11	187–350	external	x	x	x		x		x
12	350–873	NaCl+H_2O				x			
13	873–2090	NaCl+H_2O				x			
14	2090–5000	NaCl+H_2O				x			
15	350–873	internal+H_2O	x	x	x		x	x	x
16	873–2090	internal+H_2O	x		x		x		x
17	2090–5000	internal+H_2O	x		x		x		x
18	350–873	internal+H_2O			x		x		x
19	873–2090	internal+H_2O			x		x		x
20	2090–5000	internal+H_2O			x		x		x

particle inhomogeneity is taken into account in the optics model by use of the concentric core grey-shell model (Kahnert *et al.*, 2013; Andersson and Kahnert, 2016), which is likely to give a better representation of optical properties, including the backscattering coefficient, than the simple homogeneous sphere or core-shell models (e.g. Jacobson, 2000). Optical properties of pure BC particles are modelled as non-spherical fractal aggregates, using the superposition T-matrix method (Mackowski and Mishchenko, 2011). A detailed description of the optics model is given in Andersson and Kahnert (2016).

The minimization in the variational method in conjunction with a nonlinear observation operator requires a nested loop. The inner loop minimizes the cost function in Equations 19 or 26, the outer loop computes the Jacobian of the observation operator, where the derivatives are evaluated at the most recent estimate of the minimum of the cost function that was returned in the preceding run of the inner loop. The background state $\delta\widetilde{\mathbf{x}} = \mathbf{0}$ (or $\delta\mathbf{x}' = \mathbf{0}$) is employed to initialize the Jacobian in the first run of the outer loop.

The background-error covariance matrix has been modelled with a modified NMC method as described in Kahnert (2008). Horizontal error correlations were assumed to be homogeneous and isotropic. Vertical and horizontal error correlations were *not* assumed to be separable. This assumption results in vertical error correlations that are largest at intermediate horizontal wave numbers, as detailed in Kahnert (2008).

To enforce the non-negativity of the analysed mixing ratios, any instances of negative concentrations in the analysis are re-set to zero.

3.2 | Illustration of the spectral 3D-Var method with information constraints

Figure 1 illustrates the observing system simulation experiment (OSSE) of the modelling system for testing the analysis code. The MATCH model is first run with analysed meteorological input data to produce a reference run. The model output is taken as the "true" chemical state of the atmosphere. The aerosol optics model is applied to the reference aerosol field to produce "observations." More specifically, vertical profiles of the backscattering coefficient β_{bak} are computed at three typical lidar wavelengths, namely, 355, 532, and 1064 nm, as well as vertical profiles of the extinction coefficient k_{ext} at 355 and 532 nm. All five profiles are computed at a location 53.5°N, 10.0°E. The observations have not been perturbed; an observation error of 10% has been assumed. The model is run once more, but this time with 48 h forecast meteorological input data. The results of this perturbed run are taken as a proxy for a background (*a priori*) estimate of the aerosol state of the atmosphere. The "observations" and the *a priori* estimate are fed into the 3D-Var analysis code. In most tests, except the last one, only a single analysis time step is computed. The analysed state is compared to the reference state. This comparison reveals to what extent the analysis is capable of retrieving the

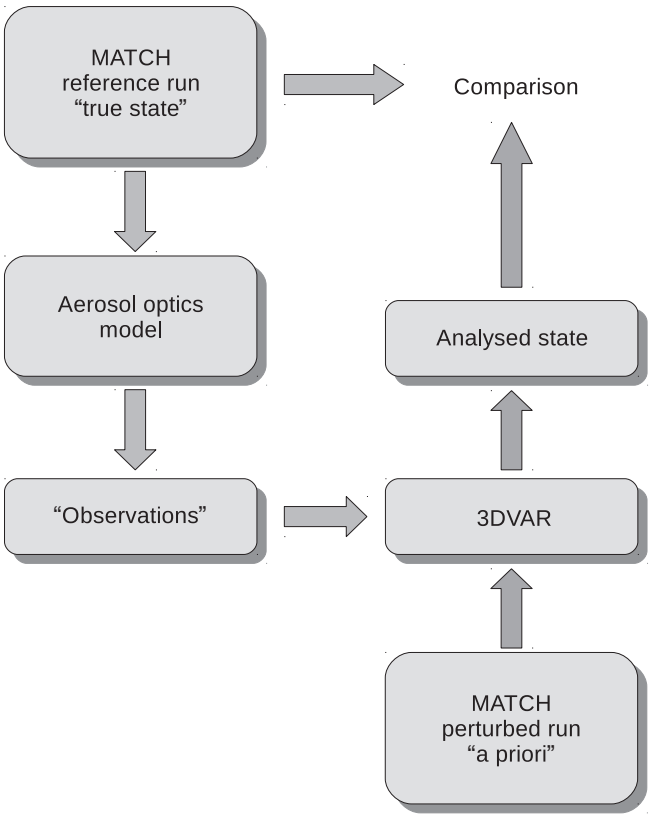

FIGURE 1 Schematic representation of the observing system simulation experiment (OSSE)

reference state by use of the observations and the *a priori* information.

The analysis step in this procedure is performed with different methods. First, the original formulation of the spectral 3D-Var method based on Equations 19 and 20 is being applied. Second, the formulation that incorporates information constraints, which is given in Equations 26 and 27, is being used. Third, one of the methods described in Wang *et al.* (2014a) has been used in one test case. In that approach the total PM_{10} mass mixing ratios $x^{i,j,l}_{\mathrm{tot}} = \sum_{k=1}^{N_c} x^{i,j,l,k}$ in each grid cell (i, j, l) are being used as control variables. The analysed state $x^{i,j,l}_{\mathrm{tot,a}}$ is distributed among the various aerosol species according to their background ratios, i.e.

$$x^{i,j,l,k}_{\mathrm{a}} = x^{i,j,l}_{\mathrm{tot,a}} \frac{x^{i,j,l,k}_{\mathrm{b}}}{x^{i,j,l}_{\mathrm{tot,b}}}. \tag{29}$$

Figure 2 shows vertical profiles of several aerosol variables at the observation site. The different panels show elemental carbon (EC, a), organic carbon (OC, b), dust (c), sea salt (NaCl, d) (each summed over all size bins), the sum of all secondary inorganic aerosols (SIA, e), as well as the total aerosol mass mixing ratio (PM_{10}, f). The reference and *a priori* profiles are indicated in black and green, respectively. The unconstrained analysis results are shown in blue, and the corresponding analysis with the information constraints is shown in red. The analysis based on the method in Wang *et al.* (2014a), in which PM_{10} is used as the only aerosol control variable, is represented by the cyan lines. Both the constrained and the unconstrained analysis results agree reasonably well. This is expected, as long as the background-error covariances are well described by our method for modelling the **B**-matrix. However, for organic carbon (top right), which in this case has a relatively low concentration, we do observe some conspicuous oscillations in the vertical direction. Most likely, the method employed for modelling the background-error covariance matrix (Kahnert, 2008 gives details) underestimates the magnitude of the vertical error correlations at large wave numbers (i.e. small horizontal length-scales), which results in insufficient vertical smoothing of the analysis results. Thus the vertical oscillations are not an inherent problem of the analysis method as such, but a peculiarity of the vertical background-error correlations that have been used here.

In the analysis that follows the method by Wang *et al.* (2014a) (cyan lines), only the total PM_{10} mass mixing ratio acts as control variable. Each aerosol component is assumed to contribute to the analysed PM_{10} mass mixing ratio in the same proportion as in the background estimate. Not surprisingly, this method yields, for most aerosol components, analysis results that reproduce the reference profiles less faithfully than either the constrained or unconstrained analyses. However, as we will see shortly, this method is much faster than the other two analysis methods.

Figure 3 shows results analogous to those of Figure 2, but 200 km north of the observation site. As expected, the analysis yields a retrieval of the reference profiles that is less faithful than the one at the observation site. However, the analysis yields, overall, still a clear improvement of the background estimate, at least for the dominant aerosol components and for PM_{10}.

In this example the total number of observations (backscattering and extinction at five wavelengths at different altitudes) is 107. The constrained analysis was performed by truncating the control vector in Equation 22 to $N_{\mathrm{cut}} = 65$ components. The corresponding singular value is $h'_{65} = 0.87$. The question is how the analysis depends on the choice of this dimensional truncation index. Figure 4 shows vertical profiles, where the panels are as in Figure 2. The background and reference values are, as before, shown in green and black, respectively. The constrained analysis results are shown for the full-size control vector, $N_{\mathrm{cut}} = 107$, and those with a truncated control vector for $N_{\mathrm{cut}} = 65$, 20, and 10. The results for $N_{\mathrm{cut}} = 107$ and 65 are almost indistinguishable. However, truncating the dimension at 20 or even 10 significantly changes (and degrades) the analysis.

A similar test has been performed with the MATCH-SALSA model, in which aerosol microphysical processes are switched on. In this case the control vector has 76 components per grid cell. The observation operator in this model version is nonlinear. Figure 5 shows the analysis results compared with the background and reference results. The first two rows show size-integrated results for (a) EC, (b) OC, (c) dust, (d) NaCl, (e) SIA, and (f) PM_{10}. (g)–(i) show the

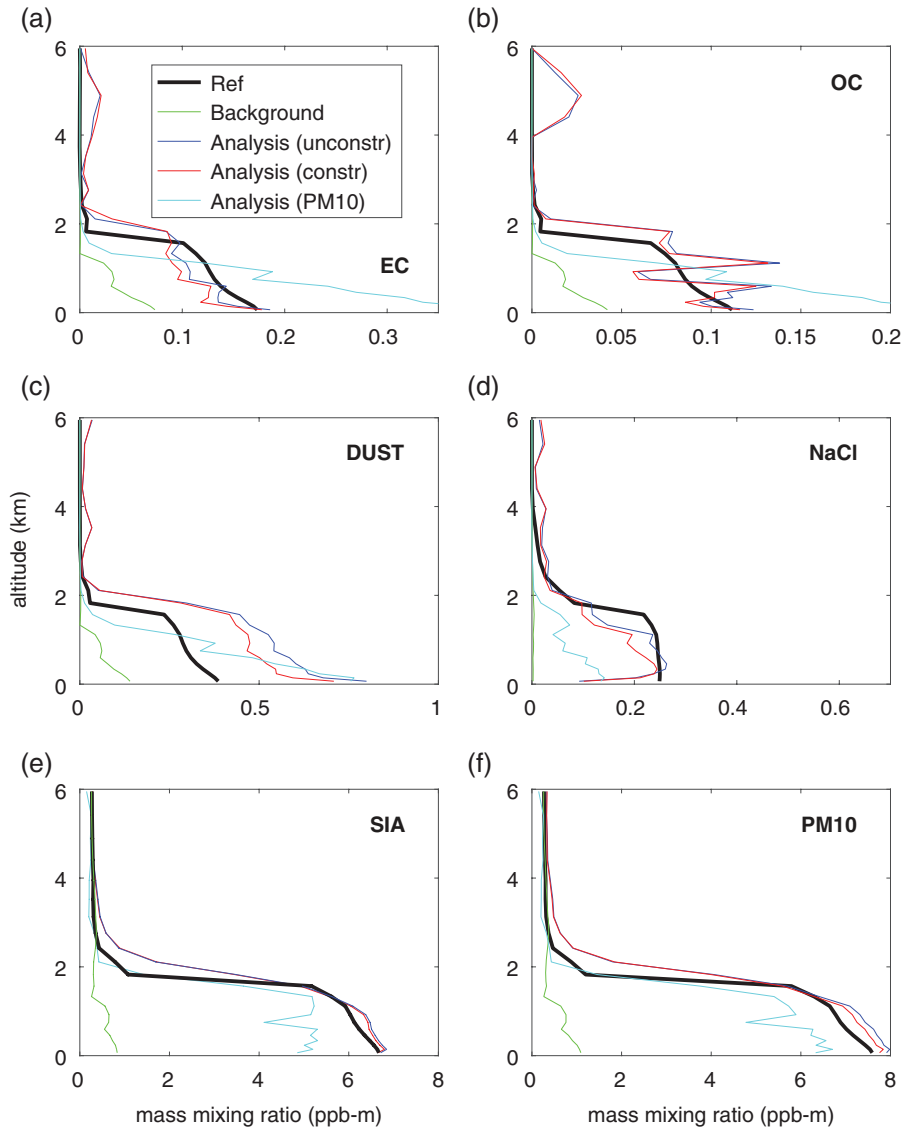

FIGURE 2 Vertical profiles of aerosol components (a) elemental carbon, EC, (b) organic carbon, OC, (c) dust, (d) sea salt, NaCl (each summed over all size bins), (e) the sum of all secondary inorganic aerosols, SIA, and (f) the total aerosol mass mixing ratio, PM_{10}. Results are shown at the observation site. The reference results (black) and the background (green) are compared to the unconstrained analysis (blue), the constrained analysis (red), and the analysis based on using PM_{10} as control variable (cyan)

sum over all aerosol components in the particle radius intervals 1.5–25 nm, 25–350 nm, and 350–5000 nm. In most cases, except OC, the analysis improves the background estimate. PM_{10}, EC and dust are retrieved quite faithfully, followed by SIA. The constrained and unconstrained analyses yield similar results. For OC the unconstrained analysis lies somewhat closer to the reference, while for NaCl and SIA the constrained analysis is slightly better, but the differences are not dramatic. For particles with a radius smaller than 25 nm (g), the analysis lies very close to the background. For particles with radii between 25 and 350 nm (h), the analysis lies close to the reference. For particles with radii larger than 350 nm (i), the analysis lies approximately in the middle between the reference and the background. This illustrates that the observations contain most information on the concentration of particles at intermediate sizes, less information on coarser particles, and almost no information on very small particles.

For the nonlinear observation operator in conjunction with the MATCH-SALSA model, the OSSE has been extended by running the analysis over the period of one month with 6 h time steps. The results are shown in Figure 6 for an altitude of 300 m at the observation site. The constrained and unconstrained analyses are almost indistinguishable. In almost all cases the analysis improves the temporal correlation with the reference in comparison to the background. This is particularly evident for SIA, PM_{10}, and for particle radii between 25 and 350 nm.

A more quantitative comparison is given in Table 2, which shows the correlation coefficients of the background and reference results, $r(\mathbf{x}_b, \mathbf{x}_{ref})$, the unconstrained analysis and the reference, $r(\mathbf{x}_{a,u}, \mathbf{x}_{ref})$, as well as the constrained analysis and the reference, $r(\mathbf{x}_{a,c}, \mathbf{x}_{ref})$. The coefficients were evaluated over the one-month time series at the observation site at an altitude of 300 m. For almost all components the

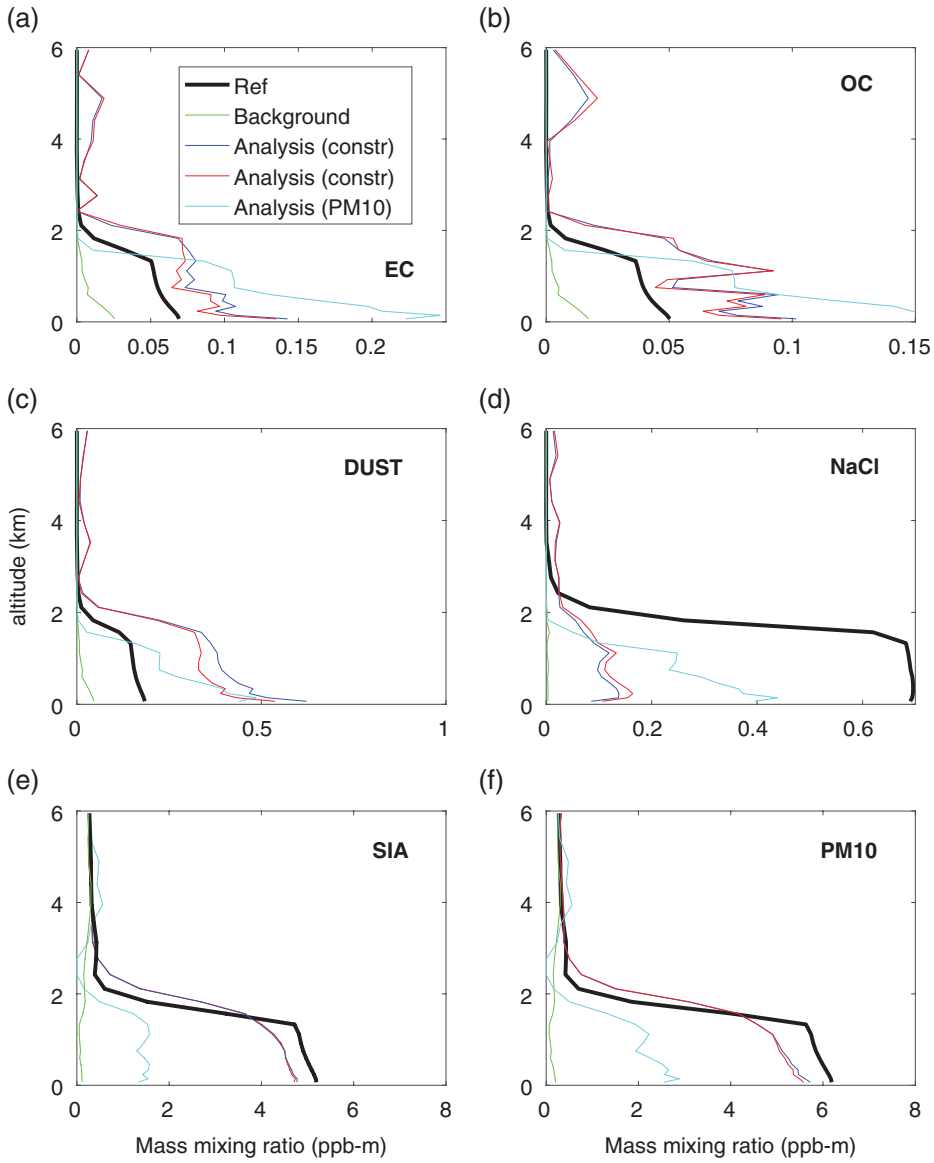

FIGURE 3 As Figure 2, but for 200 km north of the observation site

analysis achieves a significant improvement in the correlation with the reference compared with the correlation between the background and the reference; the only exception are particles in the radius range 1.5–25 nm, on which the observations carry very little information. As suspected from a visual inspection of Figure 6, the improvement is most pronounced for SIA, PM_{10}, and for particle radii between 25 and 350 nm.

Although one cannot comprehensively judge the 3D-Var formulation with information constraints based on these OSSE tests, the results obtained here are promising. The SVD approach with the constrained analysis yields, on the whole, equally good results to the conventional unconstrained analysis, as long as the truncation index N_{cut} is not chosen too small. Now the main question is whether or not the SVD approach can help us to reduce CPU time requirements. To this end, the analysis of the five lidar profiles has been performed for three different set-ups.

- First, a small domain has been considered consisting of 9×16 horizontal grid points with a resolution of 0.8°×0.8°, covering only the northern part of Germany and the southern part of Scandinavia. The mass-transport model without aerosol microphysics with 20 aerosol variables per grid point has been run on this domain. The analysis is based on using the simple linear observation operator.
- Second, the analysis was repeated with the same model version, but for a domain consisting of 83 × 105 horizontal grid points with a resolution of 0.4° × 0.4°, covering all of Europe.
- Third, the analysis was repeated for the small domain with 9 × 16 horizontal grid points, but with the SALSA microphysics model, using 76 aerosol variables per grid point. The analysis was performed by using the nonlinear observation operator. This required an extra (outer) loop for recomputing the Jacobian of the observation operator while searching for the minimum of the cost

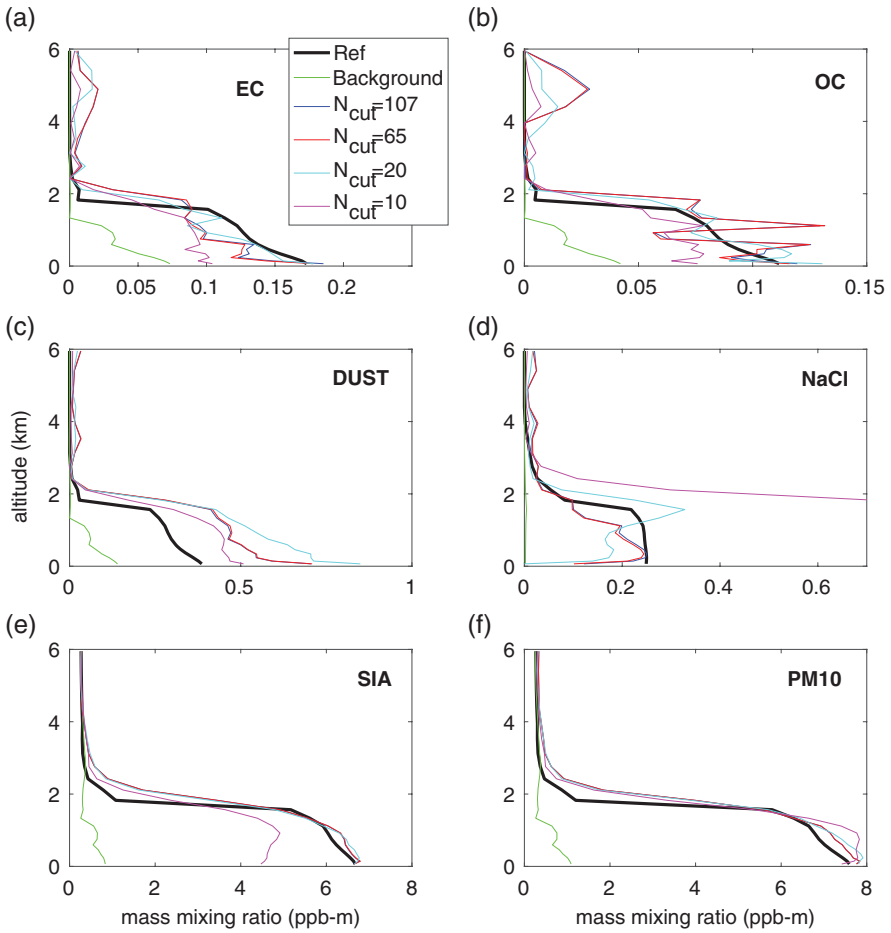

FIGURE 4 Sensitivity of the analysis result to N_{cut}, the truncation dimension of control space. The panels are as in Figure 2. The colours represent the reference (black) and background values (green), as well as the analysis with information constraints with $N_{cut} = 107$ (blue), 65 (red), 20 (cyan), and 10 (magenta). All results are shown at the observation site

function. In this test case, the outer loop went through three iterations.

In all cases the analysis was performed for 22 vertical layers.

Table 3 shows the CPU time usage for the unconstrained analysis (based on minimizing the cost function in Equation 19) and for the formulation with information constraints (based on minimizing the cost function in Equation 26). For the model version without aerosol microphysics, the required CPU time for the method by Wang *et al.* (2014a) is also shown. The code was run in serial mode. The second column from the right shows the CPU time usage for the iterative 3D-Var minimization procedure. For the SVD approach, the CPU time required for computing the SVD of the normalized Jacobian in Equation 21 is given in the rightmost column. In the constrained analysis, the CPU time usage of the 3D-Var minimization procedure is much smaller than that of the SVD. However, the CPU time usage of the SVD is, in turn, much smaller than that of the 3D-Var analysis in the conventional (unconstrained) 3D-Var minimization procedure. If we compare the CPU time usage of only the minimization algorithms in the two formulations, then the constrained formulation is about 500–1000 times faster than the conventional formulation. If we compare the sum of the CPU time usage of the 3D-Var and SVD algorithms in the constrained formulation to the required CPU time in the conventional unconstrained method, then the constrained approach is still faster by a factor of 12–16. Thus the constrained formulation entails a significant reduction in computation time. The gain in speed in the 3D-Var minimization procedure outweighs the CPU time investment required for the SVD procedure, which takes only 6–8% of the CPU time required for the minimization of the cost function in the conventional 3D-Var method.

For the test run with aerosol microphysical processes switched off, the table also shows the CPU time usage for the analysis method by Wang *et al.* (2014a), in which the PM_{10} mass mixing ratio in each grid cell is being used as control variable. This method is faster than the constrained approach by more than a factor of 10. However, as we saw earlier, it may not give equally good analysis results for individual aerosol components.

It was further found (not shown) that the choice of the cut-off index N_{cut} has only a minor effect on the CPU time usage. For instance, for the first model set-up in Table 3, the use of $N_{cut} = 65$ instead of $N_{cut} = 107$ results in a reduction

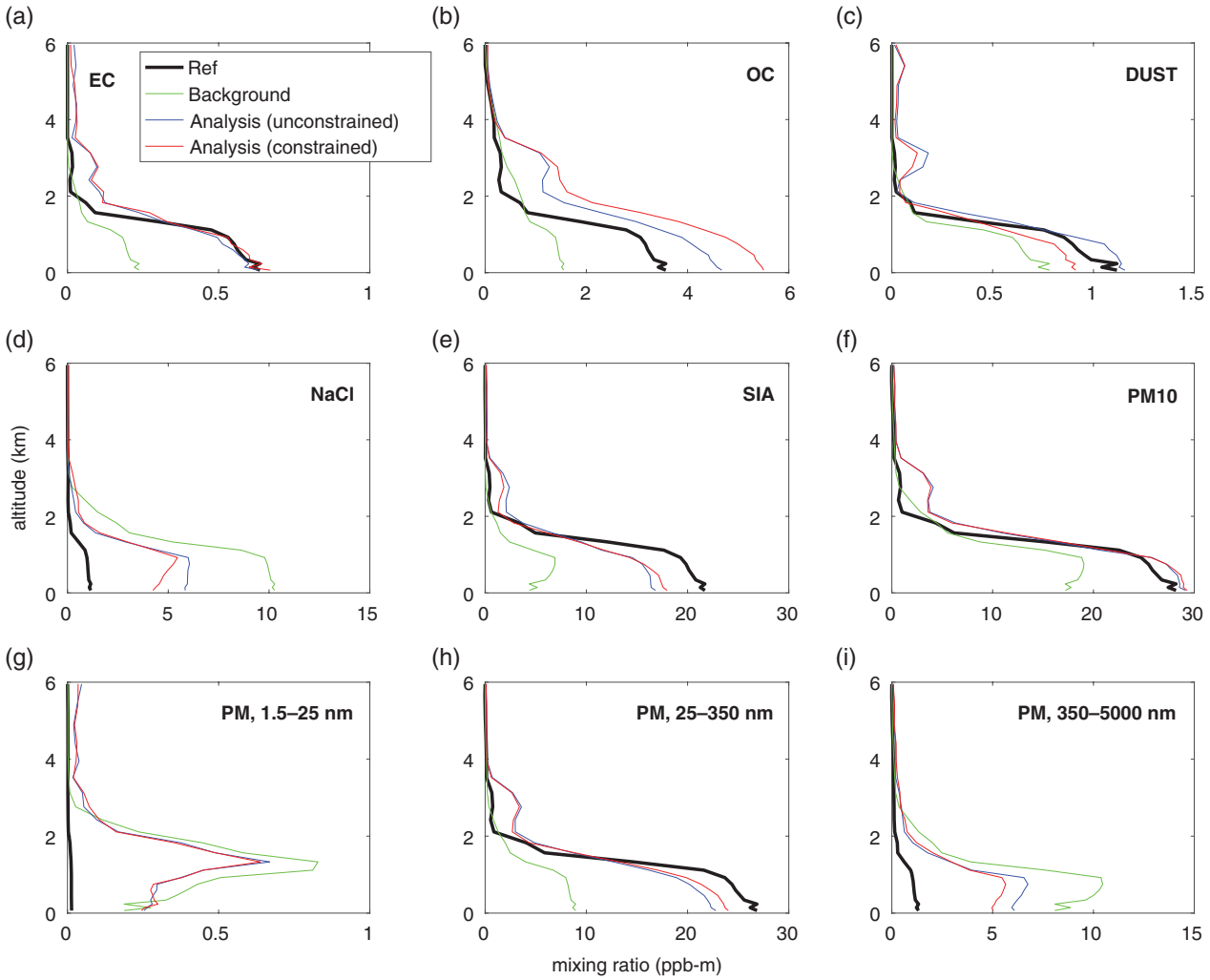

FIGURE 5 Comparison of constrained (red) and unconstrained analyses (blue) with background (green) and reference results (black). The analysis was done for the MATCH-SALSA model in conjunction with a nonlinear observation operator: mass vertical profiles of (a) EC, (b) OC, (c) dust, (d) NaCl, (e) SIA, and (f) PM_{10}, and the sum over all aerosol components in the particle-radius intervals (g) 1.5–25 nm, (h) 25–350 nm, and (i) 350–5000 nm. All profiles are shown at the observation site

of the total CPU time by 8%. Note that the choice of N_{cut} only affects the CPU time required for the 3D-Var minimization routine, but not that used by the SVD routine.

4 | DISCUSSION

The numerical tests suggest that the use of information constaints has the potential of reducing computation time in 3D-Var analyses by roughly one order of magnitude. The diagonalization of the scaled Jacobian and the reduction of the dimension of the problem from N to N_0 are the main causes for the reduction in CPU time. A further truncation of the control space from N_0 to N_{cut} results only in a minor reduction in CPU time.

Although these first results seem promising, they have to be taken with a grain of salt. First, it is as yet unclear how the method will perform in parellel computations. The present implementation makes use of the non-parallel LAPACK SVD routine SGESVD. In a future parallel version, this will have to be replaced and tested with a corresponding parallel ScaLAPACK routine. Second, in the test cases considered here, the background field differed quite significantly from the reference case. Such cases have been picked here, because it can be difficult to assess the faithfulness of an analysis in cases where the background lies already quite close to the reference. On the other hand, if the background and the reference case differ significantly, then the minimization algorithm starts far away from the minimum of the cost function, in which case a large number of iterations may be needed. This could result in a CPU time comparison that is rather favourable for the constrained approach.

Even though the information constraints confine the minimization of the cost function to a small subspace, the analysis in physical space does spread information to model variables on which the measurements contain little or no information. One way to understand this is the following. In the constrained formulation the minimization of the cost function only involves a small set of signal-related model variables. The analysis $\delta\mathbf{x}'_{\text{a}}$ is composed of this subset of variables only.

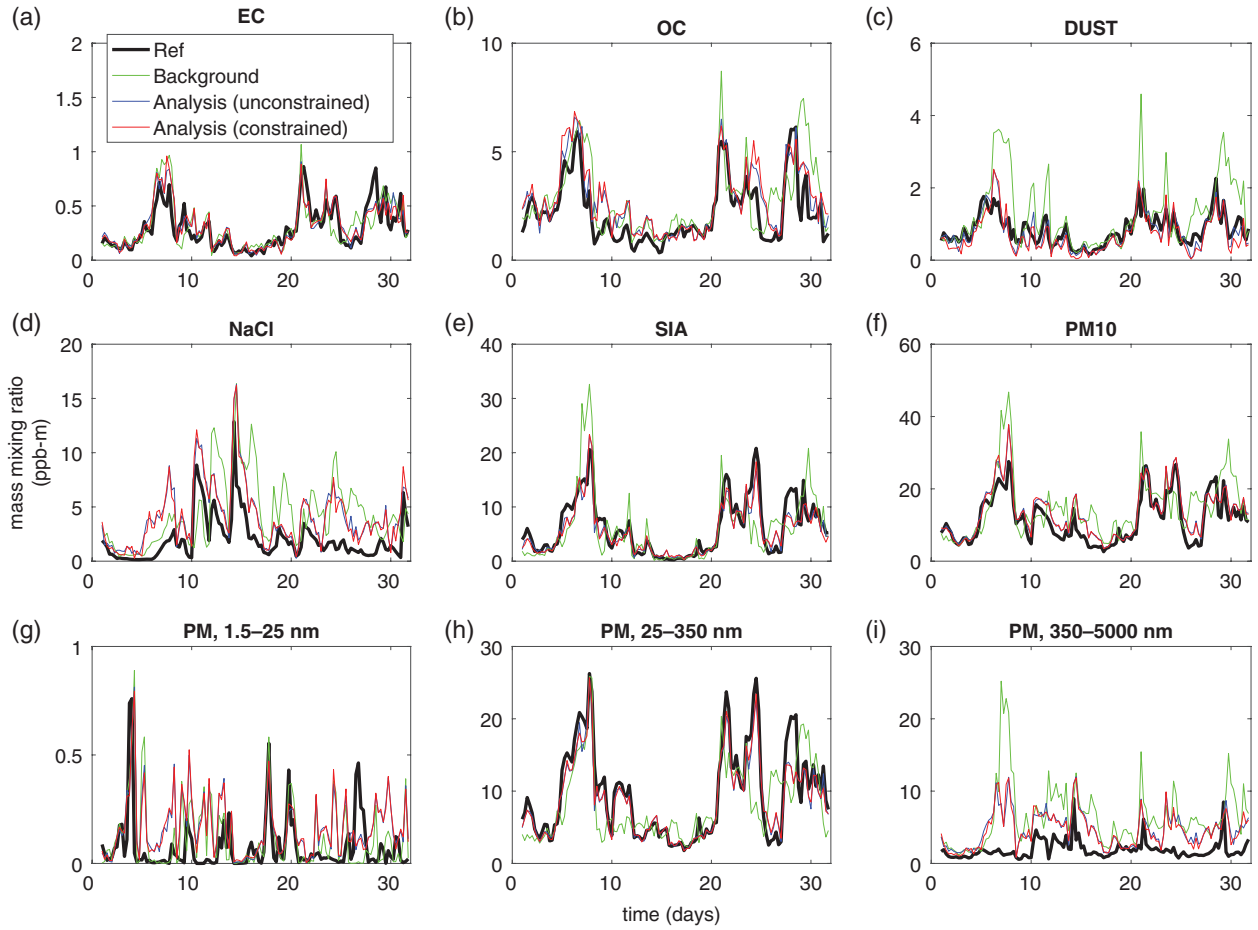

FIGURE 6 Time series of reference, background and analysis results over 1 month at an altitude of 300 m. The colours and panels are as in Figure 5

TABLE 2 Correlation coefficients of the reference results $\mathbf{x}_{\mathrm{ref}}$ with the background $\mathbf{x}_{\mathrm{b}}$, with the unconstrained analysis $\mathbf{x}_{\mathrm{a,u}}$, and with the constrained analysis $\mathbf{x}_{\mathrm{a,c}}$. Each correlation coefficient has been computed at 300 m altitude at the observation site for various aerosol components and size bins

Component	$r(\mathbf{x}_{\mathrm{b}}, \mathbf{x}_{\mathrm{ref}})$	$r(\mathbf{x}_{\mathrm{a,u}}, \mathbf{x}_{\mathrm{ref}})$	$r(\mathbf{x}_{\mathrm{a,c}}, \mathbf{x}_{\mathrm{ref}})$
EC	0.72	0.84	0.82
OC	0.67	0.89	0.87
Dust	0.60	0.86	0.82
NaCl	0.52	0.85	0.87
SIA	0.68	0.92	0.92
PM_{10}	0.62	0.93	0.92
PM 1.5–25 nm	0.35	0.33	0.33
PM 25–350 nm	0.64	0.96	0.96
PM 350–5000 nm	0.15	0.46	0.46

TABLE 3 Comparison of CPU time usage between different formulations of the 3D-Var method. "Unconstrained" refers to the case in which all aerosol components are used as independent control variables, "Constrained" refers to the method with information constraints, in which only the signal-related phase-space variables serve as control variables, and "PM_{10}" labels the approach in which the total PM_{10} mass per grid cell is taken as control variable

	Aerosol			CPU time (s)	
Grid	microphysics	Optics	Analysis	3D-Var	SVD
			Unconstrained	130	N/A
9×12	Off	Linear	Constrained	0.27	10.7
			PM_{10}	1.48	N/A
			Unconstrained	7401	N/A
83×105	Off	Linear	Constrained	7.54	455
			PM_{10}	21.5	N/A
			Unconstrained	8003	N/A
9×12	On	Nonlinear	Constrained	5.84	618

The information is spread to the rest of the model space in Equation 28 owing to the presence of the matrix $\mathbf{U}^{-1}$. By contrast, in the conventional method the information is spread throughout model space in each step of the iteration. This is because the matrix $\mathbf{U}^{-1}$ is contained in the matrix $\widetilde{\mathbf{H}}$ (Equation 18). Thus the constrained method performs the analysis in two sequential steps; the minimization is done on a small subspace, and the spreading of information is performed afterwards.

The results of this study raise the question whether or not the use of the SVD approach would have a similar effect in more advanced assimilation systems, such as 4D-Var or in an ensemble Kalman filter (EnKF).

In 4D-Var one needs to minimize a cost function

$$J(\mathbf{x}_0) = (\mathbf{x}_0 - \mathbf{x}_{\mathrm{b}})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_{\mathrm{b}}) + \sum_{i=0}^{N_w} (\mathbf{y}_i - \mathbf{H}_i \mathbf{x}_i)^{\mathrm{T}} \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{H}_i \mathbf{x}_i) \tag{30}$$

subject to the model constraint

$$\mathbf{x}_{i+1} = \mathbf{M}(t_{i+1}, t_i)\mathbf{x}_i, \qquad i = 0, \dots, N_w - 1. \tag{31}$$

Here $\mathbf{x}_0$ denotes the initial state, $\mathbf{x}_\mathrm{b}$ denotes the background estimate for the initial state, $\mathbf{B}$ represents the background-error covariance matix, and $\mathbf{y}_i$, $\mathbf{R}_i$, and $\mathbf{H}_i$ denote, respectively, the observations, observation-error covariance matrix, and the Jacobian of the observation operator at time step i. $\mathbf{M}(t_{i+1}, t_i)$ represents the tangent linear approximation of a nonlinear forward model $\mathcal{M}$. The consecutive time steps t_i, $i = 0, \dots, N_w$, cover the assimilation time window.

One can introduce a vector containing the entire sequence of observations

$$\hat{\mathbf{y}} = \left[\mathbf{y}_0^\mathrm{T}, \mathbf{y}_1^\mathrm{T}, \dots, \mathbf{y}_{N_w}^\mathrm{T}\right]^\mathrm{T}, \tag{32}$$

a block-diagonal observation-error covariance matrix

$$\hat{\mathbf{R}} = \mathrm{diag}\left\{\mathbf{R}_0, \mathbf{R}_1, \dots, \mathbf{R}_{N_w}\right\}, \tag{33}$$

and a corresponding linearized observation operator

$$\hat{\mathbf{H}} = \left[\mathbf{H}_0^\mathrm{T}, \{\mathbf{H}_1\mathbf{M}(t_1, t_0)\}^\mathrm{T} \dots, \{\mathbf{H}_{N_w}\mathbf{M}(t_{N_w}, t_0)\}^\mathrm{T}\right]^\mathrm{T}. \tag{34}$$

Then the cost function becomes

$$J(\mathbf{x}_0) = (\mathbf{x}_0 - \mathbf{x}_\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_\mathrm{b}) + (\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0)^\mathrm{T}\hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_0). \tag{35}$$

Analogous to the 3D-Var case, one can perform an SVD

$$\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{B}^{1/2} = \hat{\mathbf{V}}_\mathrm{L}\hat{\mathbf{H}}'\mathbf{V}_\mathrm{R}, \tag{36}$$

and transform the observations and the model state analogous to Equations 16, 17, 23, 24. Most likely, this would speed up the computation of the cost function and its gradient by a fraction that is comparable to that in the 3D-Var case. However, the SVD in Equation 36 would involve a matrix of higher dimension than in the corresponding 3D-Var case. Also, a subtantial fraction of the CPU time in 4D-Var is required for the forward integration of the model equations and the integration backward in time of the adjoint model. Thus the question would be how much CPU time investment the SVD in Equation 36 would require (and, indeed, if it is numerically still feasible), and how much CPU time one would save in total by diagonalizing $\hat{\mathbf{H}}$ in the evaluation of the cost function and its gradient, and in the time integration of the model and its adjoint. Since the SVD is performed outside the iteration, there may be net gain, but it is difficult to speculate on how substantial it may be. Johnson *et al.* (2005a; 2005b) did employ the SVD approach in a 4D-Var framework, but with the goal of obtaining insight into the filtering and interpolation properties of 4D-Var. The question of computational speed still needs to be investigated in a 4D-Var framework.

Kalman filter methods rely on the use of Kalman's analysis equations; for a linear observation operator $\mathbf{H}$, this reads

$$\mathbf{x}_\mathrm{a} = \mathbf{x}_\mathrm{b} + \mathbf{B}\mathbf{H}^\mathrm{T}\left[\mathbf{H}\mathbf{B}\mathbf{H}^\mathrm{T} + \mathbf{R}\right]^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}_\mathrm{b}), \tag{37}$$

where $\mathbf{x}_\mathrm{a}$ denotes the analysed state. An SVD of $\mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}^{1/2}$ in conjunction with variable transformations analogous to those in Equations 16, 17, 23, and 24 yields

$$\mathbf{x}'_\mathrm{a} = \mathbf{x}'_\mathrm{b} + \mathbf{H}'^\mathrm{T}\left[\mathbf{H}'\mathbf{H}'^\mathrm{T} + \mathbf{1}\right]^{-1}(\mathbf{y}' - \mathbf{H}'\mathbf{x}'_\mathrm{b}). \tag{38}$$

This expression reduces the products of full matrices in Equation 37 to products of the diagonal matrix $\mathbf{H}'$. By contrast to the variational method, the Kalman filter does not involve any time-consuming iterations. Therefore, it is unlikely that Equation 38 would be faster than Equation 37. On the contrary, it is quite possible that the time required for performing the SVD outweighs the gain in the matrix multiplications, resulting in a slower analysis scheme. Also, if the SVD is performed in physical space, then it may be numerically very challenging.

In the ensemble Kalman filter (Evensen, 2003) one considers a vector $\mathbf{X}^\mathrm{f} = (\mathbf{x}_1^{\mathrm{f}\mathrm{T}}, \dots, \mathbf{x}_M^{\mathrm{f}\mathrm{T}})^\mathrm{T}$ consisting of an ensemble of M model forecasts $\mathbf{x}_1^\mathrm{f}, \dots, \mathbf{x}_M^\mathrm{f}$. The error covariance matrix for the model forecast is computed from the ensemble according to

$$\mathbf{B}^\mathrm{f} = \overline{(\mathbf{X}^\mathrm{f} - \overline{\mathbf{X}^\mathrm{f}})(\mathbf{X}^\mathrm{f} - \overline{\mathbf{X}^\mathrm{f}})^\mathrm{T}}, \tag{39}$$

where the overbar denotes the ensemble mean. The analysis is performed by updating each forecast $\mathbf{x}_j^\mathrm{f}$ in the ensemble by using the Kalman analysis equation 37. Thus at each analysis time step, Equation 37 has to be applied M times. The reduction in CPU time achieved in Equation 38 can be expected to scale linearly with the ensemble size M. The SVD, on the other hand, only needs to be performed once per analysis time step. We can expect that the SVD approach, if numerically possible, results in a net reduction of CPU time for fairly large ensembles, and in a net extra cost for small ensembles.

5 | SUMMARY AND CONCLUSIONS

This work was based on applying a singular value decomposition (SVD) to the normalized Jacobian of the observation operator, which was incorporated into variational data analysis. This approach entails a transformation of the model space into a phase space, in which the model variables can be partitioned into signal- and noise-related components. One can then incorporate constraints by truncating the dimension of the phase space to restrict the minimization routine to act on the signal-related model variables only. The approach taken here differs in three essential points from the approach taken in Kahnert and Andersson (2017). First, Kahnert and Andersson (2017) performed the SVD in physical space, which required certain approximations; here, the SVD was performed in spectral space, which turned out to be numerically feasible without invoking any such approximation. Second, the approach taken here is fully Bayesian. Third, it results in a significant net reduction of CPU time. The net gain in computation time is roughly one order of magnitude. This result holds for different sizes of the spatial domain as well as for a simple 20-component and a more advanced 76-component aerosol model with a nonlinear observation operator. Numerical tests confirmed that

the dimensional truncation method yields an analysis that is close to that obtained with the conventional unconstrained approach. However, it is conceivable that this method may improve the analysis in cases where the background-error statistics are improperly specified.

The SVD approach chooses a subspace for the minimization according to the information content of the observations. By contrast, the method by Wang *et al.* (2014a) makes an educated guess that the use of PM_{10} as control variable (or of both $PM_{2.5}$ and PM_{10}) will give reasonable analysis results; it also invokes the assumption that the contribution of each aerosol component to the total aerosol mass is well represented by the background estimate. In the test case considered here, the latter assumption was rather poorly satisfied; there were differences in the mass ratios of the different aerosol components between the background and reference results. This explains why for most aerosol components the method by Wang *et al.* (2014a) gave a less faithful retrieval of the reference case than the constrained or unconstrained approach. However, the method by Wang *et al.* (2014a) is clearly superior to the constrained, and certainly to the unconstrained approach when computation time is of the essence. One can conclude that the constrained approach represents a compromise between computational speed and the faithfulness of the analysis.

The work presented here has been discussed in the context of chemical transport modelling and remote sensing of aerosol optical properties. However, the theory is rather general and not limited to any specific application of variational data analysis; it may also be useful in atmospheric chemistry, weather forecasting, ocean modelling, and other disciplines of environmental modelling and forecasting. It is conceivable that the method can also speed up 4D-Var assimilation. It is much less clear whether or not the method could reduce CPU time in an ensemble Kalman filter. If at all, a net gain can only be expected for rather large ensembles.

ACKNOWLEDGEMENTS

This work was supported by the Swedish National Space Board (Dnr. 100/16).

ORCID

Michael Kahnert https://orcid.org/0000-0001-5695-1356

REFERENCES

Andersson, E. and Kahnert, M. (2016) Coupling aerosol optics to the MATCH (v5. 5.0) chemical transport model and the SALSA (v1) aerosol microphysics module. *Geoscientific Model Development*, 9, 1803–1826.

Andersson, C., Langner, J. and Bergström, R. (2007) Interannual variation and trends in air pollution over Europe due to climate variability during 1958-2001 simulated with a regional CTM coupled to the ERA40 reanalysis. *Tellus*, 59B, 77–98.

Andersson, C., Bergström, R., Bennet, C., Robertson, L., Thomas, M., Korhonen, H., Lehtinen, K.H.J. and Kokkola, H. (2015) MATCH–SALSA – Multi-scale Atmospheric Transport and CHemistry model coupled to the SALSA aerosol microphysics model. Part 1: model description and evaluation. *Geoscientific Model Development*, 8, 171–189.

Bannister, R.N. (2017) A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143, 607–633.

Benedetti, A., Morcrette, J.-J., Boucher, O., Dethof, A., Engelen, R.J., Fisher, M., Flentje, H., Huneeus, N., Jones, L., Kinne, J.W., Mangold, A., Razinger, M., Simmons, A.J. and Suttie, M. (2009) Aerosol analysis and forecast in the European Centre for Medium-range Weather Forecasts Integrated Forecast System: 2. Data assimilation. *Journal of Geophysical Research*, 114(D13), 205.

Bocquet, M. (2009) Toward optimal choices of control space representation for geophysical data assimilation. *Monthly Weather Review*, 137, 2331–2348.

Bocquet, M. and Wu, L. (2011) Bayesian design of control space for optimal assimilation of observations. II: asymptotics solution. *Quarterly Journal of the Royal Meteorological Society*, 137, 1357–1368. https://doi.org/10.1002/qj.841..

Bocquet, M., Wu, L. and Chevallier, F. (2011) Bayesian design of control space for optimal assimilation of observations. I: consistent multiscale formalism. *Quarterly Journal of the Royal Meteorological Society*, 137, 1340–1356. https://doi.org/10.1002/qj.837.

Burton, S.P., Chemyakin, E., Liu, X., Knobelspiesse, K., Stamnes, S., Sawamura, P., Moore, R.H., Hostetler, C.A. and Ferrare, R.A. (2016) Information content and sensitivity of the $3\beta+2\alpha$ lidar measurement system for aerosol microphysical retrievals. *Atmospheric Measurement Techniques*, 9, 5555–5574.

Cardinali, C., Pezzulli, S. and Andersson, E. (2004) Influence-matrix diagnostic of a data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 130, 2767–2786.

Chen, D., Liu, Z., Schwartz, CS., Lin, HC., Cetola, JD., Gu, Y. and Xue, L. (2014) The impact of aerosol optical depth assimilation on aerosol forecasts and radiative effects during a wild fire event over the United States. *Geoscientific Model Development*, 7, 2709–2715.

Evensen, G. (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 343–367.

Evensen, G. (2007) *Data Assimilation – The Ensemble Kalman Filter*. Berlin: Springer.

Jacobson, MZ. (2000) A physically-based treatment of elemental carbon optics: implications for global direct forcing of aerosols. *Geophysical Research Letters*, 27, 217–220.

Janjić, T., Bormann, N., Bocquet, M., Carton, J.A., Cohn, S.E., Dance, S.L., Losa, S.N., Nichols, N.K., Potthast, R., Waller, J.A. and Weston, P. (2017) On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 0, 0–0. https://doi.org/10.1002/qj.3130.

Johnson, C., Hoskins, B.J. and Nichols, N.K. (2005a) Very large inverse problems in atmosphere and ocean modelling. *International Journal for Numerical Methods in Fluids*, 47, 759–771.

Johnson, C., Nichols, NK. and Hoskins, BJ. (2005b) A singular vector perspective of 4D-Var: filtering and interpolation. *Quarterly Journal of the Royal Meteorological Society*, 131, 1–19.

Joiner, J. and da Silva, AM. (1998) Efficient methods to assimilate remotely sensed data based on information content. *Quarterly Journal of the Royal Meteorological Society*, 124, 1669–1694.

Kahnert, M. (2008) Variational data analysis of aerosol species in a regional CTM: background-error covariance constraint and aerosol optical observation operators. *Tellus*, 60B, 753–770.

Kahnert, M. and Andersson, E. (2017) How much information do extinction and backscattering measurements contain about the chemical composition of atmospheric aerosol?. *Atmospheric Chemistry and Physics*, 17, 3423–3444.

Kahnert, M., Nousiainen, T. and Lindqvist, H. (2013) Models for integrated and differential scattering optical properties of encapsulated light-absorbing carbon aggregates. *Optics Express*, 21, 7974–7992.

Kokkola, H., Korhonen, H., Lehtinen, K.E.J., Makkonen, R., Asmi, A., Järvenoja, S., Anttila, T., Partanen, A.I., Kulmala, M., Järvinen, H., Laaksonen, A. and Kerminen, V.M. (2008) SALSA – a sectional aerosol module for large scale applications. *Atmospheric Chemistry and Physics*, 8(9), 2469–2483. https://doi.org/10.5194/acp-8-2469-2008.

Lahoz, W.A., Khattatov, B. and Ménard, R. (2010) *Data Assimilation – Making Sense of Observations*. Berlin: Springer.

Liu, Z., Liu, Q., Lin, H.C., Schwartz, C.S., Lee, Y.H. and Wang, T. (2011) Three-dimensional variational assimilation of MODIS aerosol optical depth:

implementation and application to a dust storm over East Asia. *Journal of Geophysical Research*, 116(D23), 206.

Mackowski, D.W. and Mishchenko, M.I. (2011) A multiple sphere T-matrix Fortran code for use on parallel computer clusters. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112, 2182–2192.

Pagowski, M., Liu, Z., Grell, G.A., Hu, M., Lin, H.C. and Schwartz, C.S. (2014) Implementation of aerosol assimilation in gridpoint statistical interpolation (v. 3.2) and WRF-Chem (v. 3.4.1). *Geoscientific Model Development*, 7, 1621–1627.

Rabier, F. (2005) Overview of global data assimilation developments in numerical weather prediction centres. *Quarterly Journal of the Royal Meteorological Society*, 131, 3215–3233.

Rabier, F., Fourrié, N., Chafaï, D. and Prunet, P. (2002) Channel selection methods for infrared atmospheric sounding interferometer radiances. *Quarterly Journal of the Royal Meteorological Society*, 128, 1011–1027.

Rodgers, C.D. (2000) *Inverse Methods for Atmospheric Sounding*. Singapore: World Scientific.

Rubin, J.I. and Collins, W.D. (2014) Global simulations of aerosol amount and size using MODIS observations assimilated with an ensemble Kalman filter. *Journal of Geophysical Research*, 119, 12780–12806.

Saide, P.E., Charmichael, G.R., Liu, Z., Schwartz, C.S., Lin, H.C., da Silva, A.M. and Hyer, E. (2013) Aerosol optical depth assimilation for a size-resolved sectional model: impacts of observationally constrained, multi-wavelength and fine mode retrievals on regional scale analysis and forecasts. *Atmospheric Chemistry and Physics*, 13, 10425–10444.

Saide, P.E., Kim, J., Song, C.H., Choi, M., Cheng, Y. and Charmichael, G.R. (2014) Assimilation of next generation geostationary aerosol optical depth retrievals to improve air quality simulations. *Geophysical Research Letters*, 41, 9188–9196. https://doi.org/10.1002/2014GL062089.

Sandu, A. and Chai, T. (2011) Chemical data assimilation – an overview. *Atmosphere*, 2, 426–463.

Schutgens, N.A., Miyoshi, T., Tekemura, T. and Nakajima, T. (2010) Applying an ensemble Kalman filter to the assimilation of AERONET observations in a global aerosol transport model. *Atmospheric Chemistry and Physics*, 10, 2561–2576.

Swinbank, R., Shutyaev, V. and Lahoz, W.A. (2003) *Data Assimilation for the Earth System*. Dordrecht: Kluwer Academic Publishers.

Veselovskii, I., Kolgotin, A., Griaznov, V., Müller, D., Franke, K. and Whiteman, D.N. (2004) Inversion of multiwavelength Raman lidar data for retrieval of bimodal aerosol size distribution. *Applied Optics*, 43, 1180–1195.

Veselovskii, I., Kolgotin, A., Müller, D. and Whiteman, D.N. (2005) Information content of multiwavelength lidar data with respect to microphysical particle properties derived from eigenvalue analysis. *Applied Optics*, 44, 5292–5303.

Wang, Y., Sartelet, K.N., Bocquet, M. and Chazette, P. (2013) Assimilation of ground versus lidar observations for PM_{10} forecasting. *Atmospheric Chemistry and Physics*, 13, 269–283.

Wang, Y., Sartelet, K.N., Bocquet, M. and Chazette, P. (2014a) Modelling and assimilation of lidar signals over greater Paris during the MEGAPOLI summer campaign. *Atmospheric Chemistry and Physics*, 14, 3511–3532.

Wang, Y., Sartelet, K.N., Bocquet, M., Chazette, P., Sicard, M., D'Amico, G., Léon, J.F., Alados-Arboledas, L., Amodeo, A., Augustin, P., Bach, J., Belegante, L., Binietoglou, I., Bush, X., Comerón, A., Delbarre, H., García-Vízcaino, D., Guerrero-Rascado, J.L., Hervo, M., Iarlori, M., Kokkalis, P., Lange, D., Molero, F., Montoux, N., Muñoz, A., Muñoz, C., Nicolae, D., Papayannis, A., Pappalardo, G., Preissler, J., Rizi, V., Rocadenbosch, F., Sellegri, K., Wagner, F. and Dulac, F. (2014b) Assimilation of lidar signals: application to aerosol forecasting in the western Mediterranean basin. *Atmospheric Chemistry and Physics*, 14, 12 031–12 053.

Xu, Q. (2006) Measuring information content from observations for data assimilation: relative entropy versus shannon entropy difference. *Tellus*, 59A, 198–209.

Zhang, J., Campbell, J.R., Reid, J.S., Westphal, D.L., Baker, N.L., Campbell, W.F. and Hyer, E.J. (2011) Evaluating the impact of assimilating CALIOP-derived aerosol extinction profiles on a global mass transport model. *Geophysical Research Letters*, 38(14), L14801. https://doi.org/10.1029/2011GL047737.

Zhang, J., Campbell, J.R., Hyer, E.J., Reid, J.S., Westphal, D.L. and Johnson, R.S. (2014) Evaluating the impact of multisensor data assimilation on a global aerosol particle transport model. *Journal of Geophysical Research*, 119, 4674–4689.

How to cite this article: Kahnert M. Information constraints in variational data assimilation. *Q J R Meteorol Soc.* 2018;144:2230–2244. https://doi.org/10.1002/qj.3347