Variational data assimilation via sparse regularisation

By ARDESHIR M. EBTEHAJ[1,2], MILIJA ZUPANSKI[3], GILAD LERMAN[2] and
EFI FOUFOULA-GEORGIOU[1]*, [1]*Department of Civil Engineering, Saint Anthony Falls Laboratory,
University of Minnesota, Minneapolis, MN, USA*; [2]*School of Mathematics, University of Minnesota,
Minneapolis, MN, USA*; [3]*Cooperative Institute for Research in the Atmosphere, Colorado State University,
Fort Collins, CO, USA*

ABSTRACT

This paper studies the role of sparse regularisation in a properly chosen basis for variational data assimilation
(VDA) problems. Specifically, it focuses on data assimilation of noisy and down-sampled observations while
the state variable of interest exhibits sparsity in the real or transform domains. We show that in the presence
of sparsity, the ℓ_1-norm regularisation produces more accurate and stable solutions than the classic VDA
methods. We recast the VDA problem under the ℓ_1-norm regularisation into a constrained quadratic
programming problem and propose an efficient gradient-based approach, suitable for large-dimensional
systems. The proof of concept is examined via assimilation experiments in the wavelet and spectral domain
using the linear advection–diffusion equation.

Keywords: variational data assimilation, sparsity, l_1 regularisation, wavelet, discrete cosine transform

1. Introduction

Environmental prediction models are initial value problems
and their forecast skills highly depend on the quality of their
initialisation. Data assimilation (DA) seeks the best esti-
mate of the initial condition of a (numerical) model, given
observations and physical constraints coming from the
underlying dynamics (see, Daley, 1993; Kalnay, 2003).
This important problem is typically addressed by two
major classes of methodologies, namely sequential and
variational methods (Ide et al., 1997; Law and Stuart,
2012). The sequential methods are typically built on the
theory of mathematical filtering and recursive weighted
least-squares (WLS) (Ghil et al., 1981; Ghil, 1989; Ghil and
Malanotte-Rizzoli, 1991; Evensen, 1994a; Anderson, 2001;
Moradkhani et al., 2005; Zhou et al., 2006; Van Leeuwen,
2010, among others), while the variational methods are
mainly rooted in the theories of constrained mathemat-
ical optimisation and batch mode WLS (e.g. Sasaki,
1970; Lorenc, 1986, 1988; Courtier and Talagrand, 1990;
Zupanski, 1993, among others).

Although recent sequential methods have received a
great deal of attention, the variational methods are still

central to the operational weather forecasting systems.
Classic formulation of the variational data assimila-
tion (VDA) typically amounts to defining a (constrained)
WLS problem whose optimal solution is the best estimate
of the initial condition, the so-called *analysis* state.
This penalty function typically encodes the weighted
sum of the costs associated with the distance of the
unknown true state to the available observations and
previous model forecast, the so-called *background* state.
Indeed, the penalty function enforces the solution to be
close enough to both observations and background
state in the weighted mean squared sense, while the
weights are characterised by the observations and the
background error covariance matrices. On the other hand,
the constraints typically enforce the analysis to follow the
underlying prognostic equations in a weak or strong sense
(see, Sasaki, 1970; Daley, 1993, p. 369). Typically, when
we constrain the analysis only to the available observa-
tions and the background state at every instant of time,
the VDA problem is called 3D-Var (e.g. Lorenc, 1986;
Parrish and Derber, 1992; Lorenc et al., 2000; Kleist
et al., 2009). On the contrary, when the analysis is also
constrained to the underlying dynamics and available
observations in a window of time, the problem is called
4D-Var (e.g. Zupanski, 1993; Rabier et al., 2000; Rawlins
et al., 2007).

---

*Corresponding author.
email: efi@umn.edu

1

Inspired by the theories of smoothing spline and Kriging interpolation in geostatistics, the first signs of using regularisation in VDA trace back to the work by Wahba and Wendelberger (1980) and Lorenc (1986), where the motivation was to impose smoothness over the class of twice differentiable analysis states. More recently, Johnson et al. (2005b) argued that, in the classic VDA problem, the sum of the squared or ℓ_2-norm of the weighted background error resembles the Tikhonov regularisation (Tikhonov et al., 1977). Specifically, by the well-known connections between the Tikhonov regularisation and spectral filtering via singular value decomposition (SVD) (e.g. see Hansen, 1998; Golub et al., 1999; Hansen et al., 2006), a new insight was provided into the interpretation and stabilising role of the background state on the solution of the classic VDA problem (see, Johnson et al., 2005a). Instead of using the ℓ_2-norm of the background error, Freitag et al. (2010) and Budd et al. (2011) suggested to modify the classic VDA cost function using the sum of the absolute values or ℓ_1-norm of the weighted background error. This assumption requires to statistically suppose that the background error is heavy tailed and can be well approximated by the family of Laplace densities (e.g. Tibshirani, 1996; Lewicki and Sejnowski, 2000). For DA of sharp atmospheric fronts, Freitag et al. (2012) kept the classic VDA cost function while further proposed to regularise the analysis state by constraining the ℓ_1-norm of its first order derivative coefficients.

In this study, inspired by our previous evidence on sparsity of rainfall fields (Ebtehaj and Foufoula-Georgiou, 2011; Ebtehaj et al., 2012), we extend the previous studies (e.g. Freitag et al., 2012; Ebtehaj and Foufoula-Georgiou, 2013) in regularised variational data assimilation (RVDA) by: (a) proposing a generalised regularisation framework for assimilating low-resolution and noisy observations, while the initial state of interest exhibits sparse representation in an appropriately chosen basis; (b) demonstrating the promise of the methodology in assimilation test problems using advection–diffusion dynamics with different error structure; and (c) proposing an efficient solution method for large-scale DA problems.

The concept of sparsity plays a central role in this paper. By definition, a state of interest is sparse in a pre-selected basis, if the number of non-zero elements of its expansion coefficients in that basis (e.g. wavelet coefficients) is significantly smaller than the overall dimension of the state in the observational space. Here, we show that if sparsity in a pre-selected basis holds, this prior information can serve to improve the accuracy and stability of DA problems. To this end, using prototype studies, different initial conditions are selected, which are sparse under the wavelet and spectral discrete cosine transformation (DCT). The promise

of the ℓ_1-norm RVDA is demonstrated via assimilating down-sampled and noisy observations in a 4D-Var setting by strongly constraining the solution to the governing advection-diffusion equation. In a broader context, we delineate the roadmap and explain how we may exploit sparsity, while the underlying dynamics and observation operator might be nonlinear. Particular attention is given to explain Monte Carlo driven approaches that can incorporate a sparse prior in the context of ensemble DA.

Section 2 reviews the classic VDA problem. In Section 3, we discuss the concept of sparsity and its relationship with ℓ_1-norm regularisation in the context of VDA problems. Results of the proposed framework and comparisons with classic methods are presented in Section 4. Section 5 is devoted to conclusions and ideas for future research, mainly focusing on the use of ensemble-based approaches to address sparse promoting VDA in nonlinear dynamics. Algorithmic details and derivations are presented in Appendix.

2. Classic VDA

At the time of model initialisation t_0, the goal of DA can be stated as that of obtaining the *analysis* state as the best estimate of the true initial state, given noisy and low-resolution *observations* and the erroneous *background* state, while the analysis needs be to consistent with the underlying model dynamics. The background state in VDA is often considered to be the previous-time forecast provided by the prognostic model. By solving the VDA problem, the analysis is then being used as the initial condition of the underlying model to forecast the next time-step and so on. In the following, we assume that the unknown true state of interest at the initial time t_0 is an m-element column vector in discrete space denoted by $\mathbf{x}_0 = [x_{0,1}, \ldots, x_{0,m}]^{\mathrm{T}} \in \mathbb{R}^m$, the noisy and low-resolution observations in the time interval $[t_0, \ldots, t_k]$ are $\mathbf{y}_i \in \mathbb{R}^n$, $i = 1, \ldots, k$, where $n \ll m$. Suppose that the observations are related to the true states by the following observation model

$$\mathbf{y}_i = \mathcal{H}(\mathbf{x}_i) + \mathbf{v}_i, \qquad (1)$$

where $\mathcal{H}: \mathbb{R}^m \to \mathbb{R}^n$ denotes the nonlinear observation operator that maps the state space into the observation space, and $\mathbf{v}_i \sim \mathcal{N}(0, \mathbf{R}_i)$ is the Gaussian observation error with zero mean and covariance $\mathbf{R}_i$.

Taking into account the sequence of available observations, $\mathbf{y}_i \in \mathbb{R}^n$, $i = 0, \ldots k$, and denoting the background state and its error covariance by $\mathbf{x}_0^b \in \mathbb{R}^m$ and $\mathbf{B} \in \mathbb{R}^{m \times m}$, the 4D-Var problem amounts to obtaining

the analysis at initial time as the minimizer of the following cost function:

$$\mathcal{J}_{4D}(\mathbf{x}_0, \ \mathbf{x}_1, \ \ldots, \mathbf{x}_k) = \sum_{i=0}^{k} \left(\frac{1}{2} ||\mathbf{y}_i - \mathcal{H}(\mathbf{x}_i)||_{\mathbf{R}_i^{-1}}^2 \right) + \frac{1}{2} ||\mathbf{x}_0^b - \mathbf{x}_0||_{\mathbf{B}^{-1}}^2, \tag{2}$$

while the solution is constrained to the underlying model equation,

$$\mathbf{x}_i = \mathcal{M}_{0,i}(\mathbf{x}_0), \ i = 0, \ldots, k. \tag{3}$$

Here, $||\mathbf{x}||_{\mathbf{A}}^2 = \mathbf{x}^T \mathbf{A} \mathbf{x}$ denotes the quadratic-norm, while $\mathbf{A}$ is a positive definite matrix and the function $\mathcal{M}_{0,i} : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a nonlinear model operator that evolves the initial state in time from t_0 to t_i.

Let us define $\mathbf{M}_{0,i}$ to be the Jacobian of $\mathcal{M}_{0,i}$ and restrict our consideration only to a linear observation operator, that is $\mathcal{H}(\mathbf{x}_i) = \mathbf{H} \mathbf{x}_i$, and thus the 4D-Var cost function reduces to

$$\mathcal{J}_{4D}(\mathbf{x}_0) = \sum_{i=0}^{k} \left(\frac{1}{2} ||\mathbf{y}_i - \mathbf{H}\mathbf{M}_{0,i}\mathbf{x}_0||_{\mathbf{R}_i^{-1}}^2 \right) + \frac{1}{2} ||\mathbf{x}_0^b - \mathbf{x}_0||_{\mathbf{B}^{-1}}^2. \tag{4}$$

By defining $\mathbf{y} = \left[\mathbf{y}_0^T, \ldots, \mathbf{y}_k^T \right]^T \in \mathbb{R}^N$, where $N = n(k+1)$, $\underline{\mathbf{H}} = \left[\left(\mathbf{H}\mathbf{M}_{0,0} \right)^T, \ldots, \left(\mathbf{H}\mathbf{M}_{0,k} \right)^T \right]^T$, and

$$\underline{\mathbf{R}} = \begin{bmatrix} \mathbf{R}_0 & 0 & \cdots & 0 \\ 0 & \mathbf{R}_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{R}_k \end{bmatrix},$$

the 4D-Var problem (4) further reduces to minimisation of the following cost function:

$$\mathcal{J}_{4D}(\mathbf{x}_0) = \frac{1}{2} ||\mathbf{y} - \underline{\mathbf{H}}\mathbf{x}_0||_{\underline{\mathbf{R}}^{-1}}^2 + \frac{1}{2} ||\mathbf{x}_0^b - \mathbf{x}_0||_{\mathbf{B}^{-1}}^2. \tag{5}$$

Clearly, eq. (5) is a smooth quadratic function of the initial state of interest $\mathbf{x}_0$. Therefore, by setting the derivative to zero, it has the following analytic minimizer as the analysis state,

$$\mathbf{x}_0^a = \left(\underline{\mathbf{H}}^T \underline{\mathbf{R}}^{-1} \underline{\mathbf{H}} + \mathbf{B}^{-1} \right)^{-1} \left(\underline{\mathbf{H}}^T \underline{\mathbf{R}}^{-1} \mathbf{y} + \mathbf{B}^{-1} \mathbf{x}_0^b \right). \tag{6}$$

Throughout this study, we used Matlab built-in function pcg.m, described by Bai et al. (1987), for obtaining classic solutions of the linear 4D-Var in eq. (6).

Accordingly, it is easy to see (e.g. Daley, 1993, p. 39) that the analysis error covariance is the inverse of the Hessian of eq. (5), as follows:

$$\mathbb{E}\left[(\mathbf{x}_0 - \mathbf{x}_0^a)(\mathbf{x}_0 - \mathbf{x}_0^a)^T \right] = \left(\underline{\mathbf{H}}^T \underline{\mathbf{R}}^{-1} \underline{\mathbf{H}} + \mathbf{B}^{-1} \right)^{-1}. \tag{7}$$

It can be shown that the analysis in the above classic 4D-Var is the conditional expectation of the true state given observations and the background state. In other words, the analysis in the classic 4D-Var problem is the unbiased minimum mean squared error (MMSE) estimator of the true state (Levy, 2008, chap.4).

3. Regularised variational data assimilation

3.1. Background

As is evident, when the Hessian (i.e. $\underline{\mathbf{H}}^T \underline{\mathbf{R}}^{-1} \underline{\mathbf{H}} + \mathbf{B}^{-1}$) in the classic VDA cost function in eq. (5) is ill-conditioned, the VDA solution is likely to be unstable with large estimation uncertainty. To study the stabilising role of the background error, motivated by the well-known relationship between the Tikhonov regularisation and spectral filtering (e.g. Golub et al., 1999), Johnson et al. (2005a, b) proposed to reformulate the classic VDA problem analogous to the standard form of the Tikhonov regularisation (Tikhonov et al., 1977). Accordingly, using a change of variable $\mathbf{z}_0 = \mathbf{C}_B^{-1/2}(\mathbf{x}_0 - \mathbf{x}_0^b)$, letting $\mathbf{B} = \sigma_b^2 \mathbf{C}_B$ and $\underline{\mathbf{R}} = \sigma_r^2 \underline{\mathbf{C}}_R$, where $\mathbf{C}_B$ and $\underline{\mathbf{C}}_R$ are the correlation matrices, the classic variational cost function was proposed to be reformulated as follows:

$$\mathcal{J}_{4D}(\mathbf{z}_0) = ||\mathbf{f} - \mathbf{G}\mathbf{z}_0||_2^2 + \mu ||\mathbf{z}_0||_2^2. \tag{8}$$

where the ℓ_2-norm is $||\mathbf{x}||_2 = (\Sigma_{i=1}^m x_i^2)^{1/2}$, $\mu = \sigma_r^2 / \sigma_b^2$, $\mathbf{G} = \underline{\mathbf{C}}_R^{-1/2} \underline{\mathbf{H}} \mathbf{C}_B^{1/2}$, and $\mathbf{f} = \underline{\mathbf{C}}_R^{-1/2} \left(\mathbf{y} - \underline{\mathbf{H}} \mathbf{x}_0^b \right)$. Hence, by solving

$$\mathbf{z}_0^a = \underset{\mathbf{z}_0}{\text{argmin}} \{ \mathcal{J}_{4D}(\mathbf{z}_0) \},$$

the analysis can be obtained as, $\mathbf{x}_0^a = \mathbf{x}_0^b + \mathbf{C}_B^{1/2} \mathbf{z}_0^a$. Having the above reformulated problem, (Johnson et al., 2005a) provided new insights into the role of the background error covariance matrix on improving the condition number of the Hessian in (5), that is the ratio between its largest and the smallest singular values, and thus stability of the classic VDA problem.

To tackle DA of sharp fronts, following the above reformulation, Freitag et al. (2012) suggested to add the smoothing ℓ_1-norm regularisation as follows:

$$\mathbf{z}_0^a = \underset{\mathbf{z}_0}{\text{argmin}} \left\{ \mathcal{J}_{R4D}(\mathbf{z}_0) + \lambda \left| \left| \Phi \left(\mathbf{C}_B^{1/2} \mathbf{z}_0 + \mathbf{x}_0^b \right) \right| \right|_1 \right\}, \tag{9}$$

where the ℓ_1-norm is $||\mathbf{x}||_1 = \Sigma_{i=1}^m |x_i|$, the non-negative λ is called the regularisation parameter, and Φ is proposed to be an approximate first-order derivative operator as follows:

$$\Phi = \begin{bmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(m-1) \times m}.$$

Notice that eq. (9) is a non-smooth optimisation as the derivative of the cost function does not exist at the origin. Freitag et al. (2012) recast this problem into a quadratic programing (QP) with both equality and inequality constraints where the dimension of the proposed QP is three times larger than that of the original problem. Note that by quadratic programming we refer to minimisation or maximisation of a quadratic function with linear constraints. It is also worth noting that, the reformulations in eqs. (8) and (9) assume that the error covariance matrices are stationary (i.e. $\mathbf{B} = \sigma_b^2 \mathbf{C_B}$, $\mathbf{R} = \sigma_r^2 \mathbf{C_R}$) and the error variance is distributed uniformly across all of the problem dimension. However, without loss of generality, a covariance matrix $\mathbf{B} \in \mathbb{R}^{m \times m}$ can be decomposed as $\mathbf{B} = \mathrm{diag}(\mathbf{s}) \, \mathbf{C_B} \, \mathrm{diag}(\mathbf{s})$, where $\mathbf{s} \in \mathbb{R}^m$ is the vector of standard deviations (Barnard et al., 2000). Therefore, while one can have an advantage in stability of computation in eqs. (8) and (9), the stationarity assumptions and computations of the square roots of the error correlation matrices might be restrictive in practice.

In the subsequent sections, beyond ℓ_1 regularisation of the first order derivative coefficients, we present a generalised framework to regularise the VDA problem in a properly chosen transform domain or basis (e.g. wavelet, Fourier, DCT). The presented formulation includes smoothing ℓ_1 and ℓ_2-norm regularisation as two especial cases and does not require any explicit assumption about the stationarity of the error covariance matrices. We recast the ℓ_1-norm RVDA into a QP with lower dimension and simpler constraints compared to the presented formulation by Freitag et al. (2012). Furthermore, we introduce an efficient gradient-based optimisation method, suitable for large-scale DA problems. Some results are presented via assimilating low-resolution and noisy observations into the linear advection–diffusion equation in a 4D-Var setting.

3.2. A generalised framework to regularise variational data assimilation in transform domains

In a more general setting, to regularise the solution of the classic VDA problem, one may constrain the magnitude of the analysis in the norm sense as follows:

$$\mathbf{x}_0^a = \underset{\mathbf{x}_0}{\mathrm{argmin}}\{\mathcal{J}_{R4D}(\mathbf{x}_0)\} \text{ subject to } ||\Phi\mathbf{x}_0||_p^p \leq c, \quad (10)$$

where $c > 0$, $\Phi \in \mathbb{R}^{m \times m}$ is any appropriately chosen linear transformation, and the ℓ_p-norm is $||\mathbf{x}||_p = (\Sigma |x_i|^p)^{1/p}$ with $p > 0$. By constraining the ℓ_p-norm of the analysis, we implicitly make the solution more stable. In other words, we bound the magnitude of the analysis state and reduce the instability of the solution due to the potential ill-conditioning of the classic cost function. Using the theory

of Lagrange multipliers, the above-constrained problem can be turned into the following unconstrained one:

$$\mathbf{x}_0^a = \underset{\mathbf{x}_0}{\mathrm{argmin}}\left\{\frac{1}{2}||\underline{\mathbf{y}} - \underline{\mathbf{H}}\mathbf{x}_0||_{\underline{\mathbf{R}}^{-1}}^2 + \frac{1}{2}||\mathbf{x}_0^b - \mathbf{x}_0||_{\mathbf{B}^{-1}}^2 + \lambda||\Phi\mathbf{x}_0||_p^p\right\}.$$
$$(11)$$

where the non-negative λ is the Lagrange multiplier or regularisation parameter. As is evident, when λ tends to zero the regularised analysis tends to the classic analysis in eq. (6), while larger values are expected to produce more stable solutions but with less fidelity to the observations and background state. Therefore, in eq. (11), the regularisation parameter λ plays an important trade-off role and ensures that the magnitude of the analysis is constrained in the norm sense while keeping it sufficiently close to observations and background state. Notice that although in special cases there are some heuristic approaches to find an optimal regularisation parameter (e.g. Hansen and O'Leary, 1993; Johnson et al., 2005b), typically this parameter is selected empirically via statistical cross-validation in the problem at hand.

It is important to note that, from the probabilistic point of view, the regularised eq. (11) can be viewed as the maximum a posteriori (MAP) Bayesian estimator. Indeed, the constraint of regularisation refers to the prior knowledge about the probabilistic distribution of the state as $p(\mathbf{x}) \propto \exp\left(-\lambda||\Phi\mathbf{x}||_p^p\right)$. In other words, we implicitly assume that under the chosen transformation Φ, the state of interest can be well explained by the family of multivariate generalised Gaussian density (e.g. Nadarajah, 2005), which includes the multivariate Gaussian ($p = 2$) and Laplace ($p = 1$) densities as special cases. As is evident, because the prior term is not Gaussian, the posterior density of the above estimator does not remain in the Gaussian domain and thus characterisation of the a posteriori covariance is not straightforward in this case.

From an optimisation view point, the above RVDA problem is convex with a unique global solution (analysis) when $p \geq 1$; otherwise, it may suffer from multiple local minima. For the special case of the Gaussian prior ($p = 2$), the problem is smooth and resembles the well-known smoothing norm Tikhonov regularisation (Tikhonov et al., 1977; Hansen, 2010). However, for the case of the Laplace prior ($p = 1$), the problem is non-smooth, and it has received a great deal of attention in recent years for solving sparse ill-posed inverse problems (see Elad, 2010, and references there in). It turns out that the ℓ_1-norm regularisation promotes sparsity in the solution. In other words, using this regularisation, it is expected that the number of non-zero elements of $\Phi\mathbf{x}_0^a$ be significantly less than the observational dimension. Therefore, if we know a priori that a specific Φ projects a large number of elements of the state

variable of interest onto (near) zero values, the ℓ_1-norm is a proper choice of the regularisation term that can yield improved estimates of the analysis state (e.g. Chen et al., 1998, 2001; Candes and Tao, 2006; Elad, 2010).

In the subsequent sections, we focus on the 4D-Var problem under the ℓ_1-norm regularisation as follows:

$$\mathbf{x}_0^a = \underset{\mathbf{x}_0}{\text{argmin}}\left\{\frac{1}{2}||\underline{\mathbf{y}} - \underline{\mathbf{H}}\mathbf{x}_0||_{\underline{\mathbf{R}}^{-1}}^2 + \frac{1}{2}||\mathbf{x}_0^b - \mathbf{x}_0||_{\mathbf{B}^{-1}}^2 + \lambda||\Phi\mathbf{x}_0||_1\right\}. \tag{12}$$

It is important to note that the presented formulation in eq. (12) shares the same solution with the problem in eq. (9) while in a more general setting, it can handle non-stationary error covariance matrices and does not require additional computational cost to obtain their square roots. It is worth nothing that the ℓ_1-norm regularised 4D-Var in eq. (12) may be alternatively recast into the following form:

$$\mathbf{x}_0^a = \underset{\mathbf{x}_0}{\text{argmin}}\left\{||\Phi\mathbf{x}_0||_1\right\} \text{ subject to } \mathcal{J}_{R4D}(\mathbf{x}_0) \le c, \tag{13}$$

where $c > 0$. This problem is a quadratically constrained linear programing problem and is closely related to the original formulation of the well-known basis pursuit approach by Chen et al. (1998). In this problem formulation, the ℓ_1-norm cost function assures that we seek an analysis with sparse projection onto the subspace spanned by the chosen basis, while the constraint enforces the analysis to be sufficiently close to the available observations.

Conceptually, by adding relevant regularisation terms, we improve the stability of the VDA problem and enforce the analysis state to follow a certain regularity. Improved stability of the regularised solution comes from the fact that the regularisation term constrains the solution magnitude and prevents it from blowing up due to the possible ill-conditioning of the VDA problem. In ill-conditioned classic VDA problems, it is easy to see that the inverse of the Hessian in (7) may contain very large elements which can spoil the analysis. However, by adding a proper regularisation term and making the problem well-posed, we shrink the size of the elements of the covariance matrix and reduce the estimation error. According to the law of bias-variance trade-off, this improvement in the analysis error covariance naturally comes at the cost of introducing a small bias in the solution, whose magnitude can be kept small by proper selection of the regularisation parameter λ (e.g. Neumaier, 1998; Hansen, 2010). Regularisation may also impose a certain degree of smoothness or regularity on the analysis state. For instance, if we think of Φ as a first order derivative operator, using the smoothing ℓ_2-norm regularisation ($\lambda||\Phi\mathbf{x}_0||_2^2$), we enforce the energy of the analysis increments to be minimal, which naturally reduces the analysis variability and makes it smoother. Therefore,

using the smoothing ℓ_2-norm regularisation in a derivative space, is naturally suitable for continuous and sufficiently smooth state variables. On the other hand, for piece-wise smooth states with isolated singularities and jumps, it turns out that the use of the ℓ_1-norm regularisation ($\lambda||\Phi\mathbf{x}_0||_1$) in a derivative space is very advantageous. Using this norm in a derivative space, we implicitly constrain the total variation of the solution, which prevents imposing extra smoothness over edges and jump discontinuities.

3.2.1. Solution method via QP. Due to the separability of the ℓ_1-norm, one of the most well-known methods, often called basis pursuit (see, Chen et al., 1998; Figueiredo et al., 2007), can be used to recast the ℓ_1-norm RVDA problem in eq. (12) to a constrained quadratic programming. Here, let us assume that $\mathbf{c}_0 = \Phi\mathbf{x}_0$, where $\mathbf{x}_0$ and $\mathbf{c}_0$ are in $\mathbb{R}^m$ and split $\mathbf{c}_0$ into its positive $\mathbf{u}_0 = \max(\mathbf{c}_0, 0)$ and negative $\mathbf{v}_0 = \max(-\mathbf{c}_0, 0)$ components such that $\mathbf{c}_0 = \mathbf{u}_0 - \mathbf{v}_0$. Having this notation, we can express the ℓ_1-norm via a linear inner product operation as $||\mathbf{c}_0||_1 = \mathbf{1}_{2m}^T\mathbf{w}_0$, where $\mathbf{1}_{2m} = [1, \ldots, 1]^T \in \mathbb{R}^{2m}$ and $\mathbf{w}_0 = [\mathbf{u}_0^T, \mathbf{v}_0^T]^T$. Thus, eq. (12) can be recast as a smooth constrained QP problem on non-negative orthant as follows:

$$\underset{\mathbf{w}_0}{\text{minimize}}\left\{\frac{1}{2}\mathbf{w}_0^T\begin{bmatrix}\mathbf{Q} & -\mathbf{Q}\\-\mathbf{Q} & \mathbf{Q}\end{bmatrix}\mathbf{w}_0 + \left(\lambda\mathbf{1}_{2m} + \begin{bmatrix}\mathbf{b}\\-\mathbf{b}\end{bmatrix}\right)^T\mathbf{w}_0\right\}$$
$$\text{subject to } \mathbf{w}_0 \succcurlyeq 0, \tag{14}$$

where, $\mathbf{Q} = \Phi^{-T}\left(\underline{\mathbf{H}}^T\underline{\mathbf{R}}^{-1}\underline{\mathbf{H}} + \mathbf{B}^{-1}\right)\Phi^{-1}$, $\mathbf{b} = -\Phi^{-T}\left(\underline{\mathbf{H}}^T\underline{\mathbf{R}}^{-1}\underline{\mathbf{y}} + \mathbf{B}^{-1}\mathbf{x}_0^b\right)$, and $\mathbf{w}_0 \succcurlyeq 0$ denotes element-wise inequality.

Clearly, given the solution $\widehat{\mathbf{w}}_0$ from eq. (14), one can easily retrieve $\widehat{\mathbf{c}}_0$ and thus the analysis state is $\mathbf{x}_0^a = \Phi\widehat{\mathbf{c}}_0$.

Euclidean projection onto the constraint set of the QP problem in eq. (14) is simpler than the formulation suggested by (Freitag et al., 2012) and allows us to use efficient and convergent gradient projection methods (e.g. Bertsekas, 1976; Serafini et al., 2005; Figueiredo et al., 2007), suitable for large-scale VDA problems. The dimension of the above problem seems twice that of the original problem; however, because of the existing symmetry in this formulation, the computational burden remains at the same order as the original classic problem (see Appendix). Another important observation is that, choosing an orthogonal transformation (e.g. orthogonal wavelet, DCT, Fourier) for Φ is very advantageous computationally, as in this case $\Phi^{-1} = \Phi^T$.

It is important to note that, for the ℓ_1-norm regularisation in eq. (14), it is easy to show that the regularisation parameter is bounded as $0 < \lambda < ||\mathbf{b}||_\infty$, where the

infinity-norm is $\|\mathbf{x}\|_\infty = \max(|x_1|, \ldots, |x_m|)$. For those values of λ greater than the upper bound, clearly the analysis state in eq. (14) is the zero vector with maximum sparsity (see Appendix).

4. Examples on linear advection–diffusion equation

4.1. Problem statement

The advection–diffusion equation is a parabolic partial differential equation with a drift and has fundamental applications in various areas of applied sciences and engineering. This equation is indeed a simplified version of the general Navier–Stocks equation for a divergence-free and incompressible Newtonian fluid where the pressure gradient is negligible. In a general form, this equation for a quantity of $\mathbf{x}(s, t)$ is

$$\frac{\partial \mathbf{x}(s, t)}{\partial t} + a(s, t)\nabla \mathbf{x}(s, t) = \epsilon \nabla^2 \mathbf{x}(s, t),$$
$$\mathbf{x}(s, 0) = \mathbf{x}_0(s), \qquad (15)$$

where $a(s, t)$ represents the velocity and $\epsilon \geq 0$ denotes the viscosity constant.

The linear ($a = $ const.) and inviscid form ($\epsilon = 0$) of eq. (15) has been the subject of modelling, numerical simulation, and DA studies of advective atmospheric and oceanic flows and fluxes. For example, Lin et al. (1998) argued that the mechanism of rain-cell regeneration can be well explained by a pure advection mechanism, Jochum and Murtugudde (2006) found that Tropical Instability Waves (TIWs) need to be modelled by horizontal advection without involving any temperature mixing length. The nonlinear inviscid form (e.g. Burgers' equation) has been used in the shallow water equation and has been subject of oceanic and tidal DA studies (e.g. Bennett and McIntosh, 1982; Evensen, 1994b). The linear and viscid form ($\epsilon > 0$) has fundamental applications in modelling of atmospheric and oceanic mixing (e.g. Lanser and Verwer, 1999; Jochum and Murtugudde, 2006; Smith and Marshall, 2009, chap. 6), land-surface moisture and heat transport (e.g. Afshar and Marino, 1978; Hu and Islam, 1995; Peters-Lidard et al., 1997; Liang et al., 1999), surface water quality modelling (e.g. Chapra, 2008, chap. 8), and subsurface mass and heat transfer studies (e.g. Fetter, 1994).

Here, we restrict our consideration only to the linear form and present a series of test problems to demonstrate the effectiveness of the ℓ_1-norm RVDA in a 4D-Var setting. It is well understood that the general solution of the linear viscid form of eq. (15) relies on the principle of super-position of linear advection and diffusion. In other words, the solution at time t is obtained via shifting the initial

condition by at, followed by a convolution with the fundamental Gaussian kernel as follows:

$$\mathcal{D}(s, t) = (4\pi\epsilon t)^{-1/2} \exp\left(\frac{-|s|^2}{4\epsilon t}\right), \qquad (16)$$

where the standard deviation is $\sqrt{2\epsilon t}$. As is evident, the linear shift of size at also amounts to obtaining the convolution of the initial condition with a Kronecker delta function as follows:

$$\mathcal{A}(s - at) = \begin{cases} 1 & s = at \\ 0 & \text{otherwise} \end{cases}. \qquad (17)$$

4.2. Assimilation set-up and results

4.2.1. Prognostic equation and observation model. It is well understood that (circular) convolution in discrete space can be constructed as a (circulant) Toeplitz matrix–vector product (e.g. Chan and Jin, 2007). Therefore, in the context of a discrete advection–diffusion model, the temporal diffusivity and spatial linear shift of the initial condition can be expressed in a matrix form by $\mathbf{D}_{0,i}$ and $\mathbf{A}_{0,i}$, respectively. In effect, $\mathbf{D}_{0,i}$ represents a Toeplitz matrix, for which its rows are filled with discrete samples of the Gaussian Kernel in eq. (16), while the rows of $\mathbf{A}_{0,i}$ contain a properly positioned Kronecker delta function.

Thus, for our case, the underlying prognostic equation; i.e. $\mathbf{x}_i = \mathbf{M}_{0,i}\mathbf{x}_0$, may be expressed as follows:

$$\mathbf{x}_i = \mathbf{A}_{0,i}\mathbf{D}_{0,i}\mathbf{x}_0. \qquad (18)$$

In this study, the low-resolution constraints of the sensing system are modelled using a linear smoothing filter followed by a down-sampling operation. Specifically, we consider the following time-invariant linear measurement operator

$$\mathbf{H} = \frac{1}{4}\begin{bmatrix} 1111 & 0000 & \cdots & 0000 \\ 0000 & 1111 & \cdots & 0000 \\ \vdots & \vdots & \vdots & \vdots \\ 0000 & 0000 & \cdots & 1111 \end{bmatrix} \in \mathbb{R}^{n \times m}, \qquad (19)$$

which maps the higher dimensional state to a lower dimensional observation space. In effect, each observation point is then an averaged and noisy representation of the four adjacent points of the true state.

4.2.2. Initial states. To demonstrate the effectiveness of the proposed ℓ_1-norm regularisation in eq. (12), we consider four different initial conditions which exhibit sparse representation in the wavelet and DCT domains (Fig. 1). In particular, we consider: (a) a flat top-hat (FTH), which is a composition of zero-order polynomials
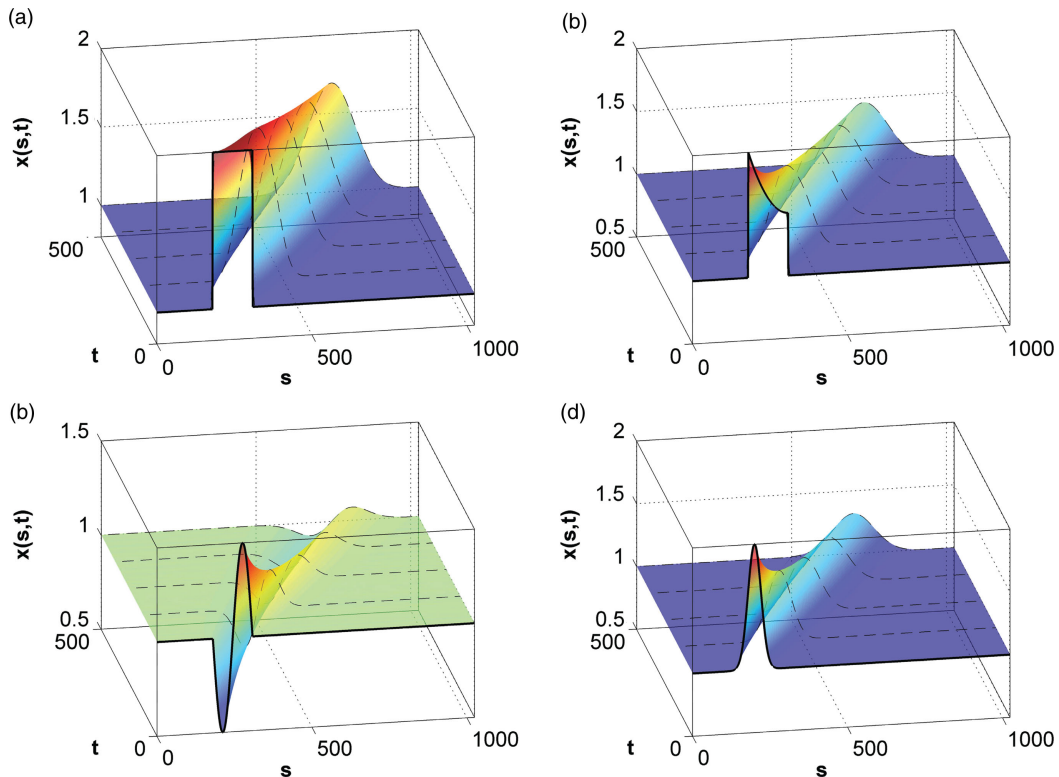
Fig. 1. Initial conditions and their evolutions with the linear advection–diffusion equation: (a) flat top-hat (FTH), (b) quadratic top-hat (QTH), (c) window sinusoid (WS), and (d) squared-exponential (SE). The first two initial conditions (a, b) exhibit sparse representation in the wavelet domain while the next two (c, d) show nearly sparse representation in the discrete cosine domain (DCT). Initial conditions are evolved under the linear advection–diffusion eq. (15) with $\epsilon = 4 \, [L^2/T]$ and $a = 1[L/T]$. The broken lines show the time instants where the low-resolution and noisy observations are available in the assimilation interval.

and can be sparsified theoretically using the first order Daubechies wavelet (DB01) or the Haar basis; (b) a quadratic top-hat which is a composition of zero and second order polynomials and theoretically can be well sparsified by wavelets with vanishing moments of order greater than three (Mallat, 2009, p. 284); (c) a window sinusoid (WS); and (d) a squared exponential function which exhibits nearly sparse behaviour in the DCT basis. All of the initial states are assumed to be in $\mathbb{R}^{1024}$ and are evolved in time with a viscosity coefficient $\epsilon = 4 \, [L^2/T]$ and velocity $a = 1[L/T]$. The assimilation interval is assumed to be between 0 and $T = 500[T]$, where the observations are sparsely available over this interval at every 125[T] time-steps (Figs. 1 and 2).

4.2.3. Observation and background error. The observations and background errors are important components of a DA system that determine the quality and information content of the analysis. Clearly, the nature and behaviour of the errors are problem-dependent and need to be

carefully investigated in a case-by-case study. It needs to be stressed that from a probabilistic point of view, the presented formulation for the ℓ_1-norm RVDA assumes that both of the error components are unimodal and can be well explained by the class of Gaussian covariance models. Here, for observation error, we only consider a stationary white Gaussian distribution, $\mathbf{v} \sim \mathcal{N}(0, \mathbf{R})$, where $\mathbf{R} = \sigma_r^2 \mathbf{I}$ (Fig. 2).

However, as discussed in (Gaspari and Cohn, 1999), the background error can often exhibit a correlation structure. In this study, the first and second order auto-regressive (AR) Gaussian Markov processes, are considered for mathematical simulation of a possible spatial correlation in the background error; see Gaspari and Cohn (1999) for a detailed discussion about the error covariance models for DA studies.

The AR(1), also known as the Ornestein–Ulenbeck process in infinite dimension, has an exponential covariance function $\rho(\tau) \propto e^{-\alpha|\tau|}$. In this covariance function, τ denotes the lag either in space or time, and the parameter α determines the decay rate of the correlation. The inverse
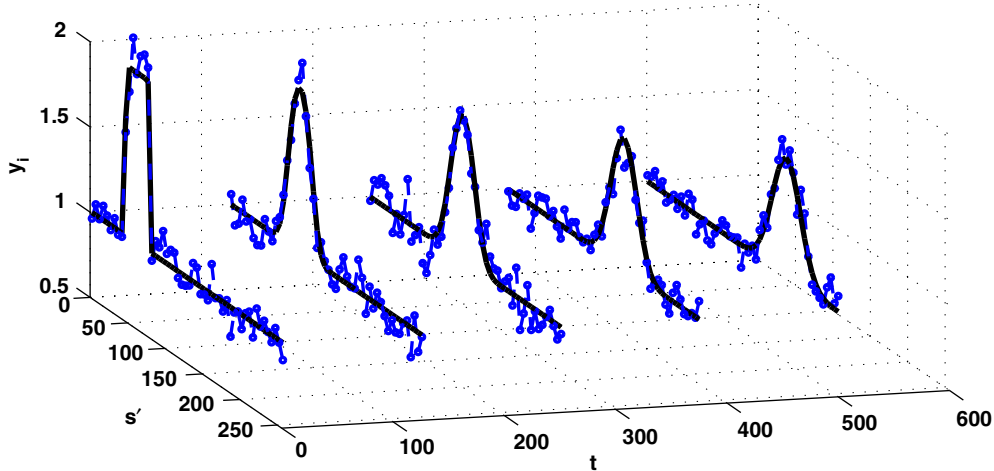
Fig. 2.   A sample representation of the available low-resolution (solid lines) and noisy observations (broken lines with circles) in every 125 [T] time-steps in the assimilation window for the flat top-hat (FTH) initial condition. Here, the observation error covariance is set to $\mathbf{R} = \sigma_r^2 \mathbf{I}$ with $\sigma_r = 0.08$ equivalent to SNR $= 20 \log \left(\sigma_{\mathbf{x}_0} / \sigma_r \right) \approx 12$ dB.

of the correlation decay rate $l_c = 1/\alpha$ is often called the characteristic correlation length of the process. The covariance function of the AR(1) model has been studied very well in the context of stochastic process (e.g. Durrett, 1999) and estimation theory (e.g. Levy, 2008). For example, it is shown by Levy (2008, p. 298) that the eigenvalues are monotonically decreasing which may give rise to a very ill-conditioned covariance matrix in the discrete space, especially for small α or large characteristic correlation lengths. The covariance function of the AR(2) is more complicated than the AR(1); however, it has been shown that in special cases, its covariance function can be explained by $\rho(\tau) \propto e^{-\alpha|\tau|}(1 + \alpha|\tau|)$ (Gaspari and Cohn, 1999; Stein, 1999, p. 31). Note that, both of these covariance models are stationary and also isotropic as they are only a function of the magnitude of the correlation lag (Rasmussen and Williams, 2006, p. 82). Consequently, the discrete background error covariance is a Hermitian Toeplitz matrix and can be decomposed into a scalar standard deviation and a correlation matrix as $\mathbf{B} = \sigma_b^2 \mathbf{C}_b$, where

$$\mathbf{C}_b = \begin{bmatrix} \rho(0) & \rho(1) & \cdots & \rho(m) \\ \rho(1) & \rho(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho(1) \\ \rho(m) & \cdots & \rho(1) & \rho(0) \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

For the same values of α, it is clear that the AR(2) correlation function decays slower than that of the AR(1). Figure 3 shows empirical estimation of the condition number of the reconstructed correlation matrices at different dimensions ranging from $m = 4$ to 1024. As is evident, the error covariance of the AR(2) has a larger condition number than that of AR(1) for the same value of

the parameter α. Clearly, as the background error plays a very important role on the overall condition number of the Hessian of the cost function in eq. (5), an ill-conditioned background error covariance makes the solution more unstable with larger uncertainty around the obtained analysis.

Figure 4 shows a sample path of the chosen error models for the background error. Generally speaking, a correlated error contains large-scale (low-frequency) components that can corrupt the main spectral components of the true state at the same frequency range. Therefore, this type of error can superimpose with the large-scale characteristic features of the initial state and its removal is naturally more difficult than that of the white error via a DA methodology.

4.3. Results of assimilation experiments

In this subsection, we present the results of the proposed regularised DA as expressed in eq. (12). We first present the results for the white background error and then discuss the correlated error scenarios. As previously explained, the first two initial conditions exhibit sharp transitions and are naturally sparse in the wavelet domain. For those initial states (Fig. 1a, b) we have used classic orthogonal wavelet transformation by Mallat (1989). Indeed, the rows of $\Phi \in \mathbb{R}^{1024 \times 1024}$ in this case contain the chosen wavelet basis that allows us to decompose the initial state of interest into its wavelet representation coefficients, as $\mathbf{c} = \Phi \mathbf{x}$ (forward wavelet transform). On the other hand, due to the orthogonality of the chosen wavelet $\Phi \Phi^T = \mathbf{I}$, columns of Φ^T contain the wavelet basis that allows us to reconstruct the initial state from its wavelet representation coefficients, that is, $\mathbf{x} = \Phi^T \mathbf{c}$ (inverse wavelet transform).
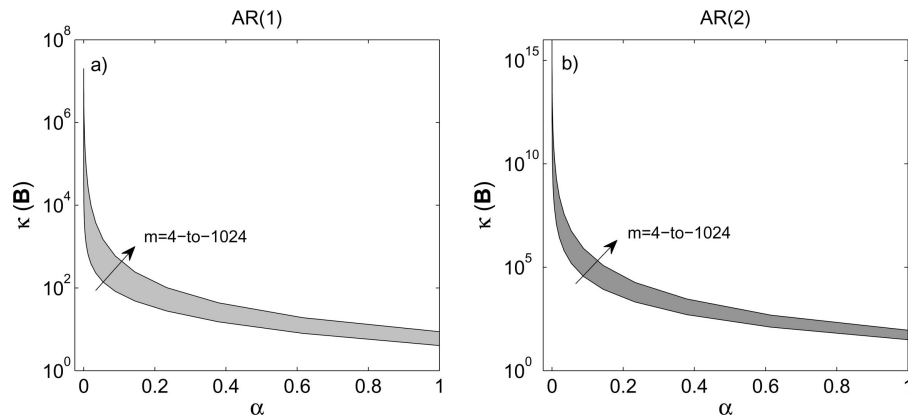
Fig. 3. Empirical condition numbers of the background error covariance matrices as a function of parameter α and problem dimension (m) for the AR(1) in (a) and AR(2) in (b). The parameter α varies along the x-axis and m varies along the different curves of the condition numbers with values between 4 and 1024. We recall that $\kappa(\mathbf{B})$ is the ratio between the largest and smallest singular values of $\mathbf{B}$. In (a) the covariance matrix is $\mathbf{B}_{ij} = e^{-\alpha|i-j|}$ and in (b) $\mathbf{B}_{ij} = e^{-\alpha|i-j|}(1 + \alpha|i-j|)$, $1 \leq i, j \leq m$. It is seen that the condition numbers of the AR(2) model are significantly larger than those of the AR(1) model for the same values of the parameter α.

We used a full level of decomposition without any truncation of wavelet decomposition levels to produce a fully sparse representation of the initial state. For example, in our case where $\mathbf{x} \in \mathbb{R}^{1024}$, we have used 10 levels of decomposition.

For the last two initial states (Fig. 1c, d), we used DCT transformation (e.g. Rao and Yip, 1990), which expresses the state of interest by a linear combination of the oscillatory cosine functions at different frequencies. It is well understood that this basis has a very strong compaction capacity to capture the energy content of sufficiently smooth states and sparsely represent them via a few elementary cosine waveforms. Note that this transformation is also orthogonal ($\Phi\Phi^{\mathrm{T}} = \mathbf{I}$) and contrary to the Fourier transformation, the expansion coefficients are real.

4.3.1. White background error. For the white background and observation error covariance matrices ($\mathbf{B} = \sigma_b^2\mathbf{I}$, $\mathbf{R} = \sigma_r^2\mathbf{I}$), we considered $\sigma_b = 0.10$ (SNR $= 20 \log(\sigma_{\mathbf{x}_0}/\sigma_r) \cong 10.5$ dB) and $\sigma_r = 0.08$ (SNR $\cong 12$ dB), respectively. Some results are shown in Fig. 5 for the selected initial conditions. It is clear that the ℓ_1-norm regularised solution markedly outperforms the classic 4D-Var solutions in terms of the selected metrics. Indeed, in the regularised analysis the error is sufficiently suppressed and filtered, while characteristic features of the initial state are well-preserved. On the other hand, classic solutions typically over-fitted and followed the background state rather than extracting the true state. As a result, we can argue that for the white error covariance the classic 4D-Var has a very weak filtering effect, which is an essential component of an ideal DA scheme. This over-fitting may be due to the redundant (over-determined)
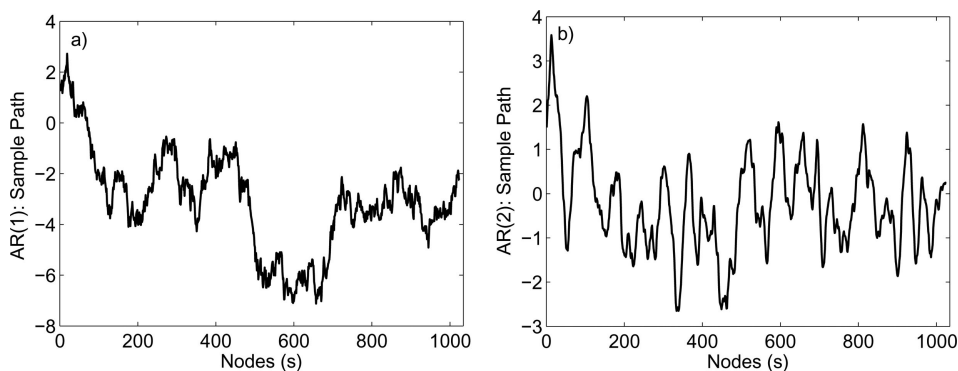

Fig. 4. Sample paths of the used correlated background error: (a) the sample path for the AR(1) covariance matrix with $\alpha^{-1} = 150$, and (b) the sample path for the AR(2) covariance matrix with $\alpha^{-1} = 25$. The paths are generated by multiplying a standard white Gaussian noise $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I})$ from the left by the lower triangular matrix $\mathbf{L}$, obtained by Cholesky factorisation of the background error covariance matrix, that is $\mathbf{B} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. It is seen that for small α, the sample paths exhibit large-scale oscillatory behaviour that can potentially corrupt low-frequency components of the underlying state.

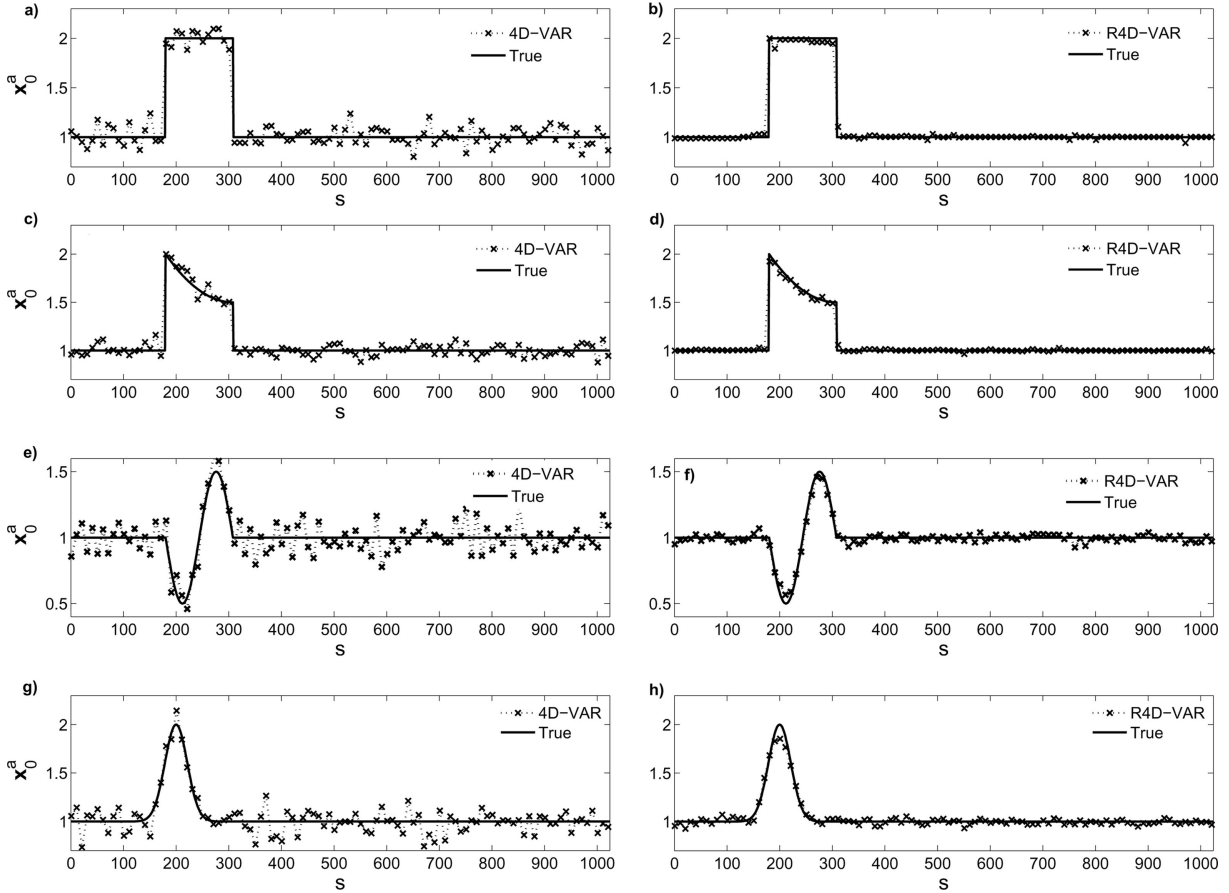*Fig. 5.* The results of the classic 4D-Var (left panel) versus the results of ℓ_1-norm R4D-Var (right panel) for the tested initial conditions in a white Gaussian error environment. The solid lines are the true initial conditions and the crosses represent the recovered initial states or the analysis. In general, the results of the classic 4D-Var suffer from overfitting while the background and observation errors are suppressed and the sharp transitions and peaks are effectively recovered in the regularised analysis.

formulation of the classic 4D-Var; see Hawkins (2004) for a general explanation on over-fitting problems in statistical estimators and also see Daley (1993, p. 41).

The average of the results for 30 independent runs is reported in Table 1. Three different lump quality metrics are examined as follows:

$$\begin{aligned} \mathrm{MSE}_r &= \left|\left|\mathbf{x}_0^t - \mathbf{x}_0^a\right|\right|_2 / \left|\left|\mathbf{x}_0^t\right|\right|_2 \\ \mathrm{MAE}_r &= \left|\left|\mathbf{x}_0^t - \mathbf{x}_0^a\right|\right|_1 / \left|\left|\mathbf{x}_0^t\right|\right|_1 \\ \mathrm{BIAS}_r &= \left|\overline{(\mathbf{x}_0^t - \mathbf{x}_0^a)}\right| / \left|\overline{\mathbf{x}}_0^t\right| \end{aligned} \quad (20)$$

namely, relative mean squared error (MSE_r), relative mean absolute error (MAE_r), and relative Bias (BIAS_r). In eq. (20), $\mathbf{x}_0^t$ denotes the true initial condition, $\mathbf{x}_0^a$ is the analysis, and overbar denotes the expected value. It is seen that based on the selected lump quality metrics, the ℓ_1-norm R4D-Var significantly outperforms the classic 4D-Var. In general, the MAE_r metric is improved more than the MSE_r metric in the presented experiments. The best improvement is obtained for the FTH initial condition, where the

sparsity is very strong compared to the other initial conditions. The MSE_r metric is improved almost three orders of magnitude, while the MAE_r improvement reaches up to six orders of magnitude in the FTH initial condition. We need to note that although the trigonometric functions can be sparsely represented in the DCT domain, here we used a WS, which suffers from discontinuities over the edges and cannot be perfectly sparsified in the DCT domain. However, we see that even in a weaker sparsity, the results of the ℓ_1-norm R4D-Var are still much better than the classic solution.

4.3.2. Correlated background error. In this part, the background error $\mathbf{B} = \sigma_b^2 \mathbf{C}_b$ is considered to be correlated. As previously discussed, typically longer correlation length creates ill-conditioning in the background error covariance matrix and makes the problem more unstable. On the other hand, the correlated background error covariance imposes

Table 1. Expected values of the MSE$_r$, MAE$_r$, and BIAS$_r$, defined in eq. (20), for 30 independent runs

	White background error					
	MSE$_r$		MAE$_r$		BIAS$_r$	
	R4D-Var	4D-Var	R4D-Var	4D-Var	R4D-Var	4D-Var
FTH	0.0188	0.0690	0.0099	0.0589	0.0016	0.0004
QTH	0.0152	0.0515	0.0083	0.0414	0.0030	0.0016
WS	0.0296	0.0959	0.0229	0.0771	0.0038	0.0022
SE	0.0316	0.0899	0.0235	0.0728	0.0018	4.26 e–5

The background and observation errors are white ($\mathbf{B} = \sigma_b^2 \mathbf{I}$, $\mathbf{R} = \sigma_r^2 \mathbf{I}$), where $\sigma_b = 0.10$ (SNR $\cong$ 10.5dB) and $\sigma_r = 0.08$ (SNR $\cong$ 12dB). The initial conditions are: flat top-hat (FTH), quadratic top-hat (QTH), window sinusoid (WS), and squared-exponential (SE). The results are reported for both the classic 4D-Var and the regularised 4D-Var (R4D-Var).

smoothness on the analysis (see, Gaspari and Cohn, 1999), improves filtering effects, and makes the classic solution to be less prone to overfitting. In this subsection, we examine the effect of correlation length on the solution of DA and compare the results of the sparsity promoting R4D-Var with the classic 4D-Var. Here, we do not apply any preconditioning as the goal is to emphasise on the stabilising role of the ℓ_1-norm regularisation in the presented formulation. In addition, for brevity, the results are only reported for the top-hat and WS initial condition, which are solved in the wavelet and DCT domains, respectively.

a)Results for the AR(1) background error

As is evident, in this case, the background state is defined by adding AR(1) correlated error to the true state (6a,d) which is known to us for these experimental studies. Figure 6 demonstrates that in the case of correlated error, the classic 4D-Var is less prone to overfitting compared to the case of the uncorrelated error in Fig. 5. Typically in the FTH initial condition with sharp transitions, the classic solution fails to capture those sharp jumps and becomes spoiled around those discontinuities (Fig. 6b). For the trigonometric initial condition (WS), the classic solution is typically overly smooth and cannot capture the peaks (Fig. 6e). These deficiencies in classic solutions typically become more pronounced for larger correlation lengths and thus more ill-conditioned problems. On the other hand, the ℓ_1-norm R4D-Var markedly outperforms the classic method by improving the recovery of the sharp transitions in FTH and peaks in WS (Fig. 6).

We examined a relatively wide range of applicable correlation lengths, $\alpha^{-1} \in \{1, 10, 25, 50, 250, 1000\}$, which correspond to the condition number $\kappa(\mathbf{B})$ of the background error covariance matrices ranging from 10^1 to 10^6 (see Fig. 3a). The assimilation results using different correlation lengths are demonstrated in Fig. 7. To have a robust conclusion about comparison of the proposed R4D-Var with the classic 4D-Var, the plots in this figure demonstrate the expected values of the quality metrics for 30 independent runs.

It can be seen that for small error correlation lengths ($\alpha^{-1} \lesssim 25$), the improvement of the R4D-Var is very significant while in the medium range ($25 \lesssim \alpha^{-1} \lesssim 50$) the classic solution becomes more competitive and closer to the regularised analysis. As previously mentioned, this improvement in the classic solutions is mainly due to the smoothing effect of the background covariance matrix. However, for larger correlation lengths ($\alpha^{-1} \gtrsim 50$), the differences of the two methods are more drastic as the classic solutions become more unstable and fail to capture the underlying structure of the initial state of interest. In general, we see that the MSE$_r$ and MAE$_r$ metrics are improved for all examined background error correlation lengths. As expected, the regularised solutions are slightly biased compared to classic solutions; however, the magnitude of the bias is not significant compared to the mean value of the initial state (see Fig. 7). Figure 7 also shows a very important outcome of regularisation which implies that the R4D-Var is almost insensitive to the studied range of correlation length and thus condition number of the problem. This confirms the stabilising role of regularisation and needs to be further studied for large-scale and operational DA problems. Another important observation is that, for extremely correlated background error, the classic 4D-Var may produce analysis with larger biases than the proposed R4D-Var (Fig. 7c). This unexpected result might be due to the presence of spurious bias in the background state coming from a strongly correlated error. In other words, a strongly correlated error may shift the mean value of the background state significantly and create a large bias in the solution of the classic 4D-Var. In this case, the improved performance of the R4D-Var may be due to its stronger stability and filtering properties.

b) Results for the AR(2) background error

The AR(2) model is suitable for errors with higher order Markovian structure compared to the AR(1) model. As is seen in Fig. (4), the condition number of the AR(2) covariance matrix is much larger than the AR(1) for the same values of the parameter α in the studied covariance
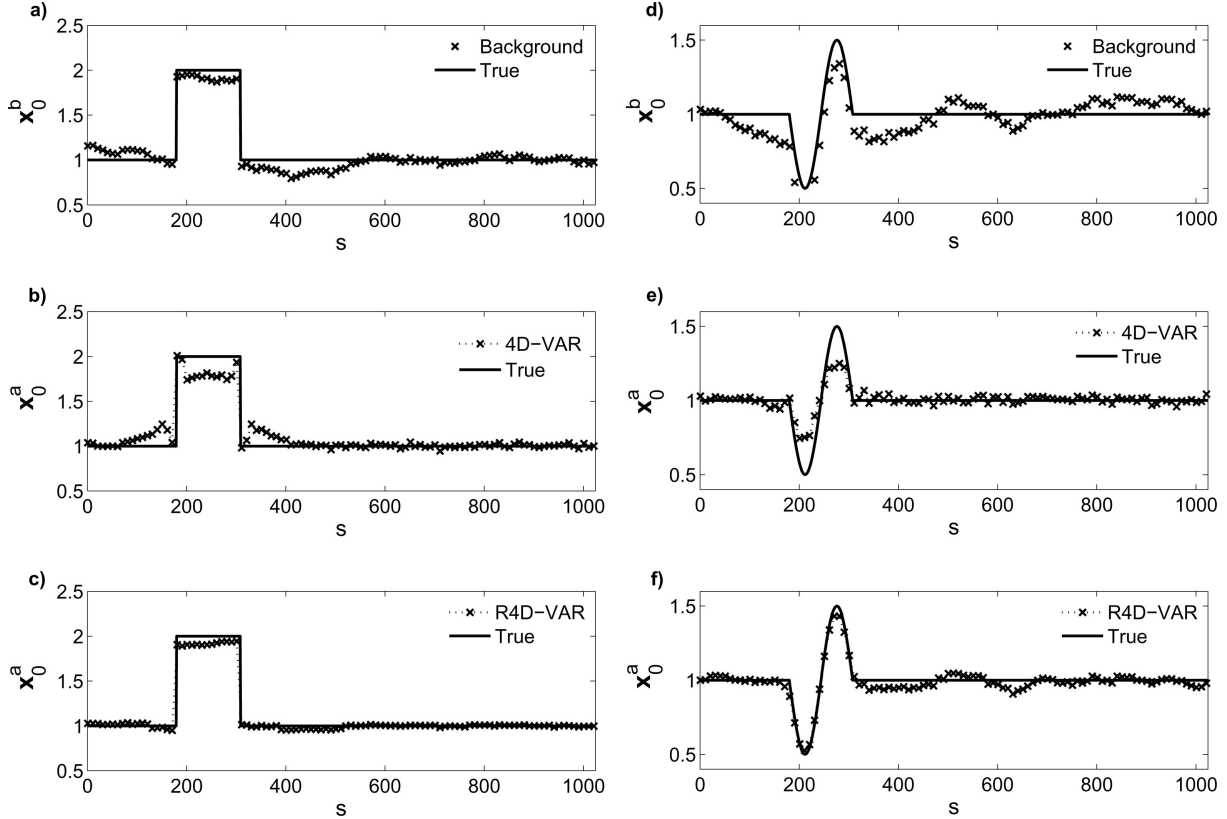
Fig. 6. Comparison of the results of the classic 4D-Var (b, ; e) and ℓ_1-norm R4D-Var (c, ; f) for the top-hat (left panel) and window sinusoid (right panel) initial conditions. The background states in (a) and (d) are defined by adding correlated errors using an AR(1) covariance model of $\rho(\tau) \propto e^{-\alpha|\tau|}$, where $\alpha = 1/250$. The results show that the ℓ_1-norm R4D-Var improves recovery of sharp jumps and peaks and results in a more stable solution compared to the classic 4D-Var; see Fig. 7 for quantitative results.

models. Here, we limited our experiments to fewer characteristic correlation lengths of $\alpha^{-1} = \{1, 5, 25, 50\}$. We constrained our considerations to $\alpha^{-1} \lesssim 50$, because for larger values (slower correlation decay rates), the condition number of **B** exceeds 10^8 and almost both methods failed to obtain the analysis without any preconditioning effort.

In our case study, for $\alpha^{-1} \lesssim 25$, where $\kappa(\mathbf{B}) \lesssim 10^6$, the proposed R4D-Var outperforms the 4D-Var similar to what has been explained for the AR(1) error in the previous subsection. However, we found that for $25 \lesssim \alpha^{-1} \lesssim 50$, where $10^6 \lesssim \kappa(\mathbf{B}) \lesssim 10^8$, without proper preconditioning, the used conjugate gradient algorithm fails to obtain the analysis state in the 4D-Var (Table 2). On the other hand, due to the role of the proposed regularisation, the R4D-Var remains sufficiently stable; however, its effectiveness deteriorated compared to the cases where the condition numbers were lower. This observation verifies the known role of the proposed regularisation for improving the condition number of the VDA problem.

4.3.3. Selection of the regularisation parameters. As previously explained, the regularisation parameter λ plays

a very important role in making the analysis sufficiently faithful to the observations and background state, while preserving the underlying regularity of the analysis. To the best of our knowledge, no general methodology exists which will produce an exact and closed form solution for the selection of this parameter, especially for the proposed ℓ_1-norm regularisation (see, Hansen, 2010, chap. 5). Here, we chose the regularisation parameter λ by trial and error based on a MMSE criterion (Fig. 8). As a rule of thumb, we found that in general $\lambda \lesssim 0.05 \|\mathbf{b}\|_\infty$ yields reasonable results. We also realised that under similar error signal-to-noise ratio, the selection of λ depends on some important factors such as, the pre-selected basis, the degree of ill-conditioning of the problem, and more importantly the ratio between the dominant frequency components of the state and the error.

5. Summary and discussion

We have discussed the concept of sparse regularisation in VDA and examined a simple but important application of the proposed problem formulation to the advection–
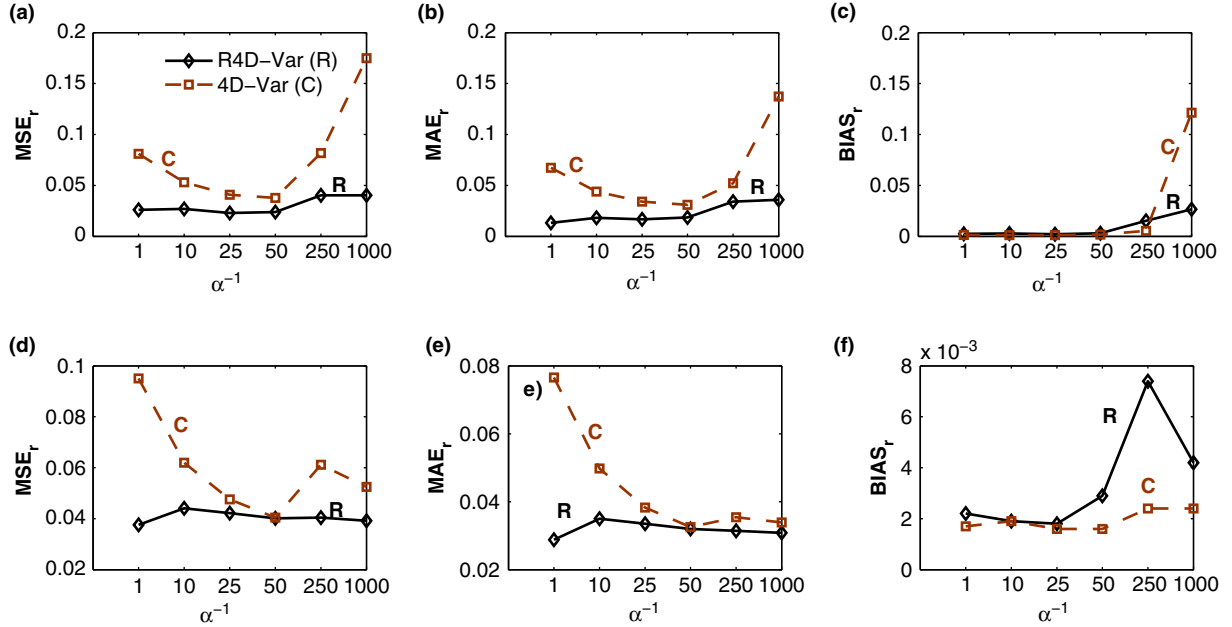
Fig. 7. Comparison of the results of the proposed ℓ_1-norm R4D-Var (solid lines) and the classic 4D-Var (broken lines) under the AR(1) background error for different correlation characteristic length scales (α^{-1}). Top panel: (a–c) the chosen quality metrics for the top-hat initial condition (FTH); Bottom panel: (d–f) the metrics for the window sinusoid initial condition (WS). These results, averaged over 30 independent runs, demonstrate significant improvements in recovering the analysis state by the proposed ℓ_1-norm R4D-Var compared to the classic 4D-Var.

diffusion equation. In particular, we extended the classic formulations by leveraging sparsity for solving DA problems in wavelet and spectral domains. The basic claim is that if the underlying state of interest exhibits sparsity in a pre-selected basis, this prior information can serve to further constrain and improve the quality of the analysis cycle and thus the forecast skill. We demonstrated that the RVDA not only shows better interpolation properties but

also exhibits improved filtering attributes by effectively removing small scale noisy features that possibly do not satisfy the underlying governing physical laws. Furthermore, it is argued that the ℓ_1-norm RVDA is more robust to the possible ill-conditioning of the DA problem and leads to more stable analysis compared to the classic methods.

We explained that, from the statistical point of view, this prior knowledge speaks for the spatial intrinsic non-

Table 2. Expected values of the MSE$_r$, MAE$_r$, and BIAS$_r$, defined in (20), for 30 independent runs

		MSE$_r$		MAE$_r$		BIAS$_r$	
	α^{-1}	R4D-Var	4D-Var	R4D-Var	4D-Var	R4D-Var	4D-Var
FTH	1	0.0254	0.0754	0.0162	0.0629	0.0023	0.0016
	5	0.0328	0.0643	0.0212	0.0534	0.0043	0.0018
	25	0.0722	–	0.0608	–	0.0187	–
	50	0.0742	–	0.0582	–	0.0268	–
WS	1	0.0363	0.0887	0.0272	0.0715	0.0029	0.0012
	5	0.0708	0.0906	0.0571	0.0529	0.0106	0.0017
	25	0.0877	–	0.0710	–	0.0243	–
	50	0.0898	–	0.0747	–	0.0361	–

AR(2) – Background error

The background and observation errors are modelled by the first order auto-regressive ($\mathbf{B} = \sigma_b^2 \mathbf{C}_B$) and white ($\mathbf{R} = \sigma_r^2 \mathbf{I}$) Gaussian processes, where $\sigma_b = 0.10$ (SNR $\cong$ 10.5dB) and $\sigma_r = 0.08$ (SNR $\cong$ 12dB). The parameter α denotes the correlation decay rate in the AR(2) covariance function $\rho(\tau) \propto e^{-\alpha|\tau|}(1 + \alpha|\tau|)$. The studied initial conditions are: flat top-hat (FTH), and window sinusoid (WS) and the results are reported for both the classic 4D-Var and the regularised 4D-Var (R4D-Var). The dash lines in the table denote that the classic method failed to return a solution without any pre-conditioning.
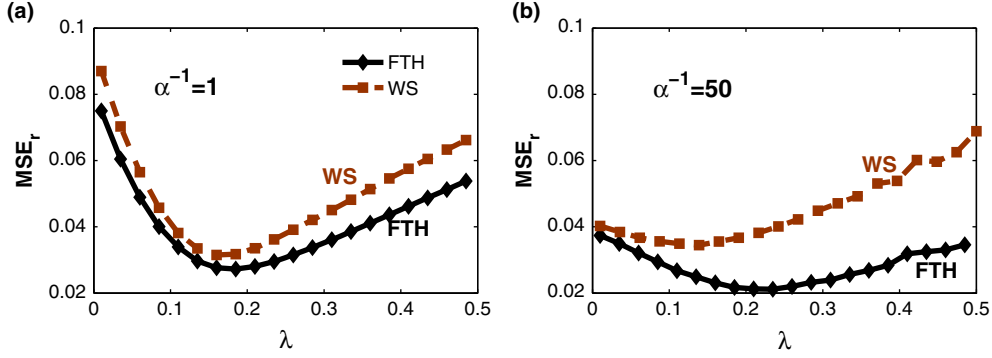
Fig. 8. The relative mean squared error versus the regularisation parameter obtained for the AR(1) background error for different characteristic correlation length (a) $\alpha^{-1}=1$, and (b) $\alpha^{-1}=50$. FTH and WS denote the flat top-hat (FTH) and window sinusoid initial conditions, respectively.

Gaussian structure of the state variable of interest, which can be well parameterised and modelled in a properly chosen basis. We discussed that selection of the sparsifying basis can be seen as a statistical model selection problem which can be guided by studying the distribution of the representation coefficients.

Note that the examined initial conditions in this study are selected under strict sparsity in the pre-selected basis, which may be compromised under realistic conditions. Additional research is required to reveal sparsity of geophysical signals in the strict and weak sense. From theoretical perspectives, further research needs to be devoted to developing methodologies to: (a) characterise the analysis covariance, especially using ensemble-based approaches; (b) automatise the selection of the regularisation parameter and study its impact on various applications of DA problems; (c) apply the methodology in an incremental setting to tackle non-linear observation operators (Courtier et al., 1994); and (d) study the role of preconditioning on the background error covariance for very ill-conditioned DA problems in RVDA settings.

Furthermore, a promising area of future research is that of developing and testing ℓ_1-norm RVDA to tackle non-linear measurement and model equations in a variational-ensemble DA setting. Basically, a crude framework can be cast as follows: (1) given the analysis and its covariance at previous time-step, properly generate an ensemble of analysis state; (2) use the analysis ensembles to generate forecasts or background ensembles via the model equation and then compute the background ensemble mean and covariance; (3) given the background ensembles, obtain observation ensembles via the observation equation and then obtain the ensemble observation covariance; (4) solve an ℓ_1-norm RVDA problem similar to that of eq. (12) for each ensemble to obtain ensemble analysis states at present time; (5) compute the ensemble analysis mean and covar-

iance and use them to forecast the next time-step; and (6) repeat the recursion.

6. Acknowledgements

7. Appendix

A.1 Quadratic Programming form of the ℓ_1-norm RVDA
To obtain the quadratic programming (QP) form presented in eq. (14), we follow the general strategy proposed in the seminal work by Chen et al. (2001). To this end, let us expand the ℓ_1-norm regularised variational data assimilation (ℓ_1-RVDA) problem in eq. (12) as follows:

$$\underset{\mathbf{x}_0}{\text{minimize}} \left\{ \frac{1}{2}\mathbf{x}_0^{\mathrm{T}}\left(\mathbf{B}^{-1} + \underline{\mathbf{H}}^{\mathrm{T}}\underline{\mathbf{R}}^{-1}\underline{\mathbf{H}}\right)\mathbf{x}_0 \right.$$
$$\left. - \left(\mathbf{B}^{-1}\mathbf{x}_0^b + \underline{\mathbf{H}}^{\mathrm{T}}\underline{\mathbf{R}}^{-1}\underline{\mathbf{y}}\right)^{\mathrm{T}}\mathbf{x}_0 + \lambda||\Phi\mathbf{x}_0||_1 \right\}. \quad (A.1)$$

Assuming $\mathbf{c}_0 = \Phi\mathbf{x}_0 \in \mathbb{R}^m$, then the above problem can be rewritten as,

$$\underset{\mathbf{c}_0}{\text{minimize}} \left\{ \frac{1}{2}\mathbf{c}_0^{\mathrm{T}}\mathbf{Q}\mathbf{c}_0 + \mathbf{b}^{\mathrm{T}}\mathbf{c}_0 + \lambda||\mathbf{c}_0||_1 \right\}, \quad (A.2)$$

where, $\mathbf{Q} = \Phi^{-\mathrm{T}}\left(\mathbf{B}^{-1} + \underline{\mathbf{H}}^{\mathrm{T}}\underline{\mathbf{R}}^{-1}\underline{\mathbf{H}}\right)\Phi^{-1}$ and $\mathbf{b} = -\Phi^{-\mathrm{T}}\left(\mathbf{B}^{-1}\mathbf{x}_0^b + \underline{\mathbf{H}}^{\mathrm{T}}\underline{\mathbf{R}}^{-1}\underline{\mathbf{y}}\right)$. Having $\mathbf{c}_0 = \mathbf{u}_0 - \mathbf{v}_0$, where $\mathbf{u}_0 = \max(\mathbf{c}_0, 0)$

$\in \mathbb{R}^m$ and $\mathbf{v}_0 = \max(-\mathbf{c}_0, 0) \in \mathbb{R}^m$ encode the positive and negative components of $\mathbf{c}_0$, problem (A.2) can be represented as follows:

$$\underset{\mathbf{u}_0, \mathbf{v}_0}{\text{minimize}} \left\{ \frac{1}{2}(\mathbf{u}_0 - \mathbf{v}_0)^\mathrm{T}\mathbf{Q}(\mathbf{u}_0 - \mathbf{v}_0) + \mathbf{b}^\mathrm{T}(\mathbf{u}_0 - \mathbf{v}_0) \right.$$
$$\left. + \lambda 1_m^\mathrm{T}(\mathbf{u}_0 + \mathbf{v}_0) \right\} \text{ subject to } \mathbf{u}_0 \succcurlyeq 0, \mathbf{v}_0 \succcurlyeq 0 \qquad \text{(A.3)}$$

Stacking $\mathbf{u}_0$ and $\mathbf{v}_0$ in $\mathbf{w}_0 = [\mathbf{u}_0^\mathrm{T}, \mathbf{v}_0^\mathrm{T}]^\mathrm{T}$, the more standard QP formulation of the problem is immediately followed as:

$$\underset{\mathbf{w}_0}{\text{minimize}} \left\{ \frac{1}{2}\mathbf{w}_0^\mathrm{T}\begin{bmatrix} \mathbf{Q} & -\mathbf{Q} \\ -\mathbf{Q} & \mathbf{Q} \end{bmatrix}\mathbf{w}_0 + \left(\lambda 1_{2m} + \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix} \right)^\mathrm{T}\mathbf{w}_0 \right\}$$
$$\text{subject to } \mathbf{w}_0 \succcurlyeq 0$$
$$\text{(A.4)}$$

Obtaining $\widehat{\mathbf{w}}_0 = [\widehat{\mathbf{u}}_0^\mathrm{T}, \widehat{\mathbf{v}}_0^\mathrm{T}]^\mathrm{T}\mathbb{R} \in^{2m}$ as the solution of (A.4), one can easily recover $\widehat{\mathbf{c}}_0 = \widehat{\mathbf{u}}_0 - \widehat{\mathbf{v}}_0$ and thus the initial state of interest $\widehat{\mathbf{x}}_0 = \Phi^{-1}\widehat{\mathbf{c}}_0$.

The dimension of the QP representation (A.4) is twice that of the original ℓ_1-RVDA problem (A.1). However, using iterative first order gradient-based methods, which are often the only practical option for large-scale data assimilation problems, it is easy to show that the effect of this dimensionality enlargement is minor on the overall cost of the problem; this is because one can easily see that obtaining the gradient of the cost function in (A.4) only requires to compute

$$\begin{bmatrix} \mathbf{Q} & -\mathbf{Q} \\ -\mathbf{Q} & \mathbf{Q} \end{bmatrix}\mathbf{w}_0 = \begin{bmatrix} \mathbf{Q}(\mathbf{u}_0 - \mathbf{v}_0) \\ -\mathbf{Q}(\mathbf{u}_0 - \mathbf{v}_0) \end{bmatrix},$$

which mainly requires matrix-vector multiplication in $\mathbb{R}^m$(e.g. Figueiredo et al., 2007).

A.2 Upper Bound of the Regularisation Parameter
Here, to derive the upper bound for the regularisation parameter in the ℓ_1-RVDA problem, we follow a similar approach as suggested for example by Kim et al. (2007). Let us refer back to the problem (A.2), which is convex but not differentiable at the origin. Obviously, $\mathbf{c}_0^a$ is a minimizer if and only if the cost function $\mathcal{J}_{R4D}(\mathbf{c}_0)$ in (A.2) is sub-differentiable at $\mathbf{c}_0^a$ and thus
$$0 \in \partial \mathcal{J}_{R4D}(\mathbf{c}_0^a),$$
where, $\partial \mathcal{J}_{R4D}(\mathbf{c}_0^a)$ denotes the sub-differential set at the solution point or analysis coefficients in the selected basis. Given that

$$\partial \mathcal{J}_{R4D}(\mathbf{c}_0^a) = \mathbf{Q}\mathbf{c}_0^a + \mathbf{b} + \lambda\partial\left(||\mathbf{c}_0^a||_1\right)$$

we have

$$-\mathbf{Q}\mathbf{c}_0^a - \mathbf{b} \in \lambda\partial\left(||\mathbf{c}_0^a||_1\right).$$

and thus for $\mathbf{c}_0^a = 0_m$, $0_m = [0, \ldots, 0]^\mathrm{T} \in \mathbb{R}^m$, one can obtain the following vector inequality

$$-\lambda 1_m \preccurlyeq -\mathbf{b} \preccurlyeq \lambda 1_m$$

which implies that $||\mathbf{b}||_\infty \leq \lambda$. Therefore λ must be less than $||\mathbf{b}||_\infty$ to obtain nonzero analysis coefficients in problem (A.2) and thus (A.1).

A.3 Gradient Projection Method
Gradient projection (GP) method is an efficient and convergent optimisation method to solve convex optimisation problems over convex sets (see, Bertsekas, 1999, p. 228). This method is of particular interest, especially, when the constraints form a convex set $\mathcal{C}$ with simple projection operator. The cost function $\mathcal{J}_{R4D}(\mathbf{w}_0)$ in eq. (14) is a quadratic function that needs to be minimised on non-negative orthant $\mathcal{C} = \{\mathbf{w}_0|\ w_{0,i} \geq 0\ \forall i = 1, \ldots, 2m\}$ as follows:

$$\widehat{\mathbf{w}}_0 = \text{argmin}\{\mathcal{J}_{R4D}(\mathbf{w}_0)\}$$
$$\text{subject to } \mathbf{w}_0 \succcurlyeq 0 \qquad \text{(A.5)}$$

For this particular problem, the GP method amounts obtained the following fixed point:

$$\mathbf{w}_0^* = [\mathbf{w}_0^* - \beta\nabla\mathcal{J}_{R4D}(\mathbf{w}_0^*)]^+, \qquad \text{(A.6)}$$

where β is a step-size along the descent direction and for every element of w_0

$$[w_0]^+ = \begin{cases} 0 & \text{if } w_0 \leq 0 \\ w_0 & \text{otherwise,} \end{cases} \qquad \text{(A.7)}$$

denotes the Euclidean projection operator onto the non-negative orthant. As is evident, the fixed point can be obtained iteratively as

$$\mathbf{w}_0^{k+1} = \left[\mathbf{w}_0^k - \beta^k\nabla\mathcal{J}_{R4D}(\mathbf{w}_0^k)\right]^+. \qquad \text{(A.8)}$$

Thus, if the descent at step k is feasible, that is $\mathbf{w}_0^k - \beta^k\nabla\mathcal{J}_{R4D}(\mathbf{w}_0^k) \succcurlyeq 0$, the GP iteration becomes an ordinary unconstrained steepest descent method, otherwise the result is mapped back onto the feasible set by the projection operator in (A.7). In effect, the GP method iteratively finds the closest feasible point in the constraint set to the solution of the original unconstrained minimisation.

In our study, the step-size β^k was selected using the *Armijo rule*, or the so-called *backtracking line search*, that is a convergent and very effective step-size rule. This step-size rule depends on two constants $0 < \xi < 0.5$, $0 < \varsigma < 1$ and is assumed to be $\beta^k = \varsigma^{m_k}$, where m_k is the smallest non-negative integer for which

$$\mathcal{J}_{R4D}\left(\mathbf{w}_0^k - \beta^k\nabla\mathcal{J}_{R4D}(\mathbf{w}_0^k)\right)$$
$$\leq \mathcal{J}_{R4D}(\mathbf{w}_0^k) - \xi\beta^k\nabla\mathcal{J}_{R4D}(\mathbf{w}_0^k)^\mathrm{T}\nabla\mathcal{J}_{R4D}(\mathbf{w}_0^k). \qquad \text{(A.9)}$$

A closer look at this line search scheme shows that it begins with a unit step-size in the direction of the negative of the gradient and reduces it by the parameter ς until the stopping criterion in (A.9) is met. In our experiments, the backtracking parameters are set to $\xi = 0.2$ and $\varsigma = 0.5$ (see, Boyd and Vandenberghe, 2004, pp. 464 for further explanation). In our coding, the iterations terminate if $\frac{||\mathbf{w}_0^k - \mathbf{w}_0^{k-1}||_2}{||\mathbf{w}_0^{k-1}||_2} \leq 10^{-5}$ or if the number of iterations exceeds 100.

References

Afshar, A. and Marino, M. A. 1978. Model for simulating soil-water content considering evapotranspiration. *J. Hydrol.* **37**, 309–322.

Anderson, J. L. 2001. An ensemble adjustment Kalman Filter for data assimilation. *Mon. Weather Rev.* **129**, 2884–2903.

Bai, Z., Demmel, J., Dongarra, J., Ruhe, A. and Van Der Vorst, H. 1987. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide* Vol. 11, SIAM, Philadelphia.

Barnard, J., McCulloch, R. and Meng, X. 2000. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sin.* **10**, 1281–1312.

Bennett, A. F. and McIntosh, P. C. 1982. Open ocean modeling as an inverse problem: tidal theory. *J. Phys. Oceanogr.* **12**, 1004–1018.

Bertsekas, D. 1976. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Automat. Contr.* **21**, 174–184.

Bertsekas, D. P. 1999. *Nonlinear Programming*. 2nd ed. Athena Scientific, Belmont, MA, 794 pp.

Boyd, S. and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press, New York, 716 pp.

Budd, C., Freitag, M. and Nichols, N. 2011. Regularization techniques for ill-posed inverse problems in data assimilation. *Comput. Fluids*. **46**, 168–173.

Candes, E. and Tao, T. 2006. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theor.* **52**, 5406–5425.

Chan, R. H.-F. and Jin, X.-Q. 2007. *An Introduction to Iterative Toeplitz Solvers*. SIAM, Philadelphia.

Chapra, S. C. 2008. *Surface Water Quality Modeling*. Waveland Press, Inc., long grove, IL, USA.

Chen, S., Donoho, D. and Saunders, M. 2001. Atomic decomposition by basis pursuit. *SIAM Rev.* **43**, 129–159.

Chen, S. S., Donoho, D. and Saunders, M. 1998. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**, 33–61.

Courtier, P. and Talagrand, O. 1990. Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus A*. **42**, 531–549.

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. 1994. A strategy for operational implementation of 4D-VAR, using an incremental approach. *Q. J. Roy. Meteorol. Soc.* **120**, 1367–1387.

Daley, R. 1993. *Atmospheric Data Analysis*. Cambridge University Press, New York, NY, USA. 472 pp.

Durrett, R. 1999. *Essentials of Stochastic Processes*. Springer-Verlag, New York Inc., New York, NY, USA.

Ebtehaj, A. M. and Foufoula-Georgiou, E. 2011. Statistics of precipitation reflectivity images and cascade of Gaussian-scale mixtures in the wavelet domain: a formalism for reproducing extremes and coherent multiscale structures. *J. Geophys. Res.* **116**, D14110.

Ebtehaj, A. M. and Foufoula-Georgiou, E. 2013. On variational downscaling, fusion and assimilation of hydro-meteorological states: a unified framework via regularization. **49**, 5944–5963.

Ebtehaj, A. M., Foufoula-Georgiou, E. and Lerman, G. 2012. Sparse regularization for precipitation downscaling. *J. Geophys. Res.* **116**, D22110.

Elad, M. 2010. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer-Verlag, New York Inc., New York, NY, USA. 376 pp.

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

Evensen, G. 1994b. Inverse methods and data assimilation in nonlinear ocean models. *Physica D*. **77**, 108–129.

Fetter, C. 1994. *Applied Hydrogeology*. 4th ed. Prentice Hall, New Jersey.

Figueiredo, M., Nowak, R. and Wright, S. 2007. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Sel. Topics Signal Process.* **1**, 586–597.

Freitag, M. A., Nichols, N. K. and Budd, C. J. 2010. L1-regularisation for ill-posed problems in variational data assimilation. *PAMM*. **10**, 665–668.

Freitag, M. A., Nichols, N. K. and Budd, C. J. 2012. Resolution of sharp fronts in the presence of model error in variational data assimilation. *Q. J. Roy. Meteorol. Soc.* **139**, 742–757.

Gaspari, G. and Cohn, S. E. 1999. Construction of correlation functions in two and three dimensions. *Q. J. Roy. Meteorol. Soc.* **125**, 723–757.

Ghil, M. 1989. Meteorological data assimilation for oceanographers. Part I: description and theoretical framework. *Dynam. Atmos. Oceans*. **13**, 171–218.

Ghil, M., Cohn, S., Tavantzis, J., Bube, K. and Isaacson, E. 1981. Applications of estimation theory to numerical weather prediction. In: *Dynamic Meteorology: Data Assimilation Methods, Applied Mathematical Sciences* (eds. L. Bengtsson, M. Ghil, and E. Källén) Vol. 36, Springer, New York, pp. 139–224.

Ghil, M. and Malanotte-Rizzoli, P. 1991. *Data Assimilation in Meteorology and Oceanography*. Elsevier Science, B.V. Amsterdam, Netherlands. pp. 141–266.

Golub, G., Hansen, P. and O'Leary, D. 1999. Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.* **21**, 185–194.

Hansen, P. 1998. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion* Vol 4. Society for Industrial Mathematics (SIAM), Philadelphia.

Hansen, P. 2010. *Discrete Inverse Problems: Insight and Algorithms* Vol 7. Society for Industrial & Applied Mathematics (SIAM), Philadelphia, PA.

Hansen, P. and O'Leary, D. 1993. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J Sci Comput*. **14**, 1487–1503.

Hansen, P., Nagy, J. and O'Leary, D. 2006. *Deblurring Images: Matrices, Spectra, and Filtering* Vol 3. Society for Industrial & Applied Mathematics (SIAM), Philadelphia, PA.

Hawkins, D. M. 2004. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12.

Hu, Z. and Islam, S. 1995. Prediction of ground surface temperature and soil moisture content by the force–restore method. *Water Resour. Res.* **31**, 2531–2539.

Ide, K., Courtier, P., Gill, M. and Lorenc, A. 1997. Unified notation for data assimilation: operational, sequential, and variational. *J. Met. Soc. Japan.* **75**, 181–189.

Jochum, M. and Murtugudde, R. 2006. Temperature advection by tropical instability waves. *J. Phys. Oceanogr.* **36**, 592–605.

Johnson, C., Hoskins, B. J. and Nichols, N. K. 2005a. A singular vector perspective of 4D-Var: filtering and interpolation. *Q. J. Roy. Meteorol. Soc.* **131**, 1–19.

Johnson, C., Nichols, N. K. and Hoskins, B. J. 2005b. Very large inverse problems in atmosphere and ocean modelling. *Int. J. Numer. Meth. Fluids.* **47**, 759–771.

Kalnay, E. 2003. *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge University Press, New York, 341 pp.

Kim, S.-J., Koh, K., Lustig, M., Boyd, S. and Gorinevsky, D. 2007. An interior-point method for large-scale l1-regularized least squares. *IEEE J. Sel. Topics Signal Process.* **1**, 606–617.

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Wu, W. S. and co-authors. 2009. Introduction of the GSI into the NCEP global data assimilation system. *Weather. Forecast.* **24**, 1691–1705.

Lanser, D. and Verwer, J. 1999. Analysis of operator splitting for advection–diffusion–reaction problems from air pollution modelling. *J. Comput. Appl. Math.* **111**, 201–216.

Law, K. J. H. and Stuart, A. M. 2012. Evaluating data assimilation algorithms. *Mon. Weather. Rev.* **140**(11), 3757–3782.

Levy, B. C. 2008. *Principles of Signal Detection and Parameter Estimation*. 1st ed. Springer Publishing Company, New York, 639 pp.

Lewicki, M. and Sejnowski, T. 2000. Learning overcomplete representations. *Neural Comput.* **12**, 337–365.

Liang, X., Wood, E. F. and Lettenmaier, D. P. 1999. Modeling ground heat flux in land surface parameterization schemes. *J. Geophys. Res.* **104**, 9581–9600.

Lin, Y.-L., Deal, R. L. and Kulie, M. S. 1998. Mechanisms of cell regeneration, development, and propagation within a two-dimensional multicell storm. *J. Atmos. Sci.* **55**, 1867–1886.

Lorenc, A. 1988. Optimal nonlinear objective analysis. *Q. J. Roy. Meteorol. Soc.* **114**, 205–240.

Lorenc, A. C. 1986. Analysis methods for numerical weather prediction. *Q. J. Roy. Meteorol. Soc.* **112**, 1177–1194.

Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingleby, N. B., Andrews, P. L. F. and co-authors. 2000. The Met. Office global three-dimensional variational data assimilation scheme. *Q. J. Roy. Meteorol. Soc* **126**, 2991–3012.

Mallat, S. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674–693.

Mallat, S. 2009. *A Wavelet Tour of Signal Processing: The Sparse Way*. 3rd ed. Elsevier, 805 pp.

Moradkhani, H., Hsu, K.-L., Gupta, H. and Sorooshian, S. 2005. Uncertainty assessment of hydrologic model states and parameters: sequential data assimilation using the particle filter. *Water Resour. Res.* **41**, W05012.

Nadarajah, S. 2005. A generalized normal distribution. *J. Appl. Stat.* **32**, 685–694.

Neumaier, A. 1998. Solving Ill-conditioned and singular linear systems: a tutorial on regularization. *SIAM Rev.* **40**, 636–666.

Parrish, D. F. and Derber, J. C. 1992. The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Weather Rev.* **120**, 1747–1763.

Peters-Lidard, C. D., Zion, M. S. and Wood, E. F. 1997. A soil–vegetation–atmosphere transfer scheme for modeling spatially variable water and energy balance processes. *J. Geophys. Res.* **102**, 4303–4324.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F. and Simmons, A. 2000. The ECMWF operational implementation of four-dimensional variational assimilation. I: experimental results with simplified physics. *Q. J. Roy. Meteorol. Soc.* **126**, 1143–1170.

Rao, K. and Yip, P. 1990. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press, Boston.

Rasmussen, C. and Williams, C. 2006. *Gaussian Processes for Machine Learning* Vol 1, MIT press, Cambridge, MA.

Rawlins, F., Ballard, S. P., Bovis, K. J., Clayton, A. M., Li, D. and co-authors. 2007. The Met Office global four-dimensional variational data assimilation scheme. *Q. J. Roy. Meteorol. Soc.* **133**, 347–362.

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

Serafini, T., Zanghirati, G. and Zanni, L. 2005. Gradient projection methods for quadratic programs and applications in training support vector machines. *Optim. Methods Softw.* **20**, 353–378.

Smith, K. S. and Marshall, J. 2009. Evidence for enhanced Eddy mixing at middepth in the Southern ocean. *J. Phys. Oceanogr.* **39**, 50–69.

Stein, M. L. 1999. *Interpolation of Spatial Data*. Springer-Verlag, New York.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288.

Tikhonov, A., Arsenin, V. and John, F. 1977. *Solutions of Ill-Posed Problems*. Winston & Sons, Washington, DC.

Van Leeuwen, P. J. 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Q. J. Roy. Meteorol. Soc.* **136**, 1991–1999.

Wahba, G. and Wendelberger, J. 1980. Some new mathematical methods for variational objective analysis using splines and cross validation. *Mon. Weather. Rev.* **108**, 1122–1143.

Zhou, Y., McLaughlin, D. and Entekhabi, D. 2006. Assessing the performance of the ensemble Kalman filter for land surface data assimilation. *Mon. Weather Rev.* **134**, 2128–2142.

Zupanski, M. 1993. Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon. Weather Rev.* **121**, 2396–2408.