School of Mathematical
and Physical Sciences

University of

Department of Mathematics and Statistics

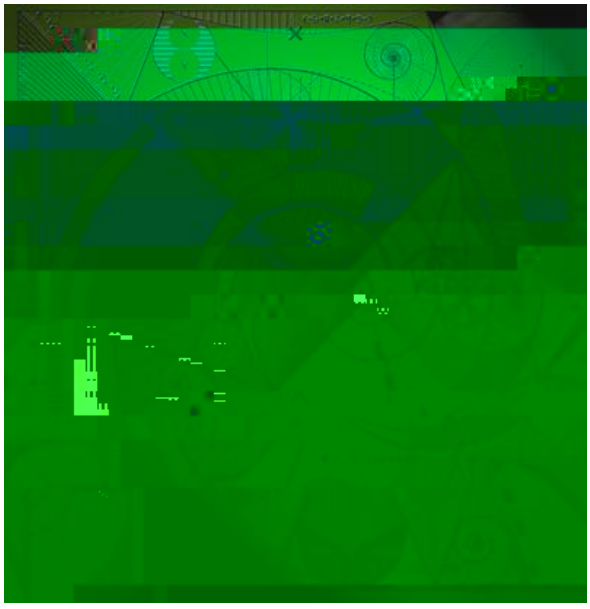
Preprint MPS-2012-18

3 September 2012

Variational data assimilation for very large environmental problems

by

A.S. Lawless



Variational data assimilation for very large environmental problems

A.S. Lawless

School of Mathematical and Physical Sciences, University of Reading, PO Box 220, Whiteknights, Reading, RG6 6AX
a.s.lawless@reading.ac.uk

Abstract

Variational data assimilation is commonly used in environmental forecasting to estimate the current state of the system from a model forecast and observational data. The assimilation problem can be written simply in the form of a nonlinear least squares optimization problem. However the practical solution of the problem in large systems requires many careful choices to be made in the implementation. In this article we present the theory of variational data assimilation and then discuss in detail how it is implemented in practice. Current solutions and open questions are discussed.

KEYWORDS: 3D-Var, 4D-Var, adjoint model, background errors, error covariance, incremental formulation, nested models, observation errors, optimization, reduced order models, tangent linear model, weak-constraint.

1 Introduction

Data assimilation is the process of combining a numerical model forecast with observational data in order to estimate the current state of a dynamical system. It has been an essential part of numerical weather prediction (NWP) since its beginnings in the 1940s, when it was recognized that errors in the initial model state could rapidly lead to large errors in the forecast. Early data assimilation schemes were based on a simple interpolation between the observations and the model state, with later schemes also taking account of the statistics of the errors in the data. Such schemes included smoothing splines, successive correction, optimal interpolation and analysis correction [68], [71]. The possible use of methods based on variational calculus was proposed by Sasaki [86], [87]

The author is supported in part by the U.K. Natural Environment Research Council, through the National Centre for Earth Observation.

in the late 1950s and 1960s, but at the time a practical implementation was not possible. A real breakthrough in the application of variational schemes to NWP came in the late 1980s with a series of papers demonstrating how the

that are related to the model state through the equation

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

where $\mathcal{H}_i : \mathbb{R}^{p_i} \rightarrow \mathbb{R}^n$ is known as the observation operator and maps the state vector to observation space. The observation errors $\boldsymbol{\epsilon}_i$ are usually assumed to be unbiased, serially uncorrelated, Gaussian errors with known covariance matrices $\mathbf{R}_i$. For the numerical weather prediction problem the vector $\mathbf{x}_i$ would contain several meteorological variables, such as pressure, temperature and the three-dimensional wind, at each grid point of the model domain. The observation operator $\mathcal{H}_i$ may just be a simple interpolation in space, if the state variable is observed directly. However, it could be a much more complicated nonlinear

data assimilation is discussed in more detail in section 3.4. On each iteration of such methods the value of the cost function and its gradient at the current iterate must be calculated. The gradient of (3) with respect to the initial state $\mathbf{x}_0$ can be found by first solving the discrete adjoint equations

$$\boldsymbol{\lambda}_i = \mathbf{M}_i^T \boldsymbol{\lambda}_{i+1} - \mathbf{H}_i^T \mathbf{R}^{-1}(\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i) \tag{4}$$

where $\boldsymbol{\lambda}_i$ are the adjoint variables, with $\boldsymbol{\lambda}_{N+1} = 0$, and $\mathbf{H}_i$ and $\mathbf{M}_i$ are the Jacobians of the nonlinear operators $\mathcal{H}_i$ and $\mathcal{M}_i$ with respect to the state variable $\mathbf{x}_i$. In the data assimilation literature these Jacobians are referred to as the *tangent linear operator* and the *tangent linear model* (TLM). The gradient of the cost function with respect to the initial state is then given by

$$\nabla \mathcal{J}(\mathbf{x}_0) = -\boldsymbol{\lambda}_0 + \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b); \tag{5}$$

where the operators $\mathbf{H}_i^T$ and $\mathbf{M}_i^T$ are the adjoints of the observation operator and the nonlinear model. We note that these adjoints are usually taken with respect to the Euclidean inner product and therefore the adjoint is equivalent to the matrix transpose of the Jacobians. Other inner products are only necessary where a physical interpretation of the adjoint variables is required. Each iteration of a numerical optimization method therefore requires one run of the forward model (1) to calculate the value of the cost function and one run of the adjoint model (4) to calculate the gradient. This makes 4D-Var very expensive from a computational point of view.

2.1 Incremental variational data assimilation

The possibility of implementing variational data assimilation in an operational setting came with the proposal of incremental variational data assimilation [19]. In this formulation the solution to the nonlinear miminimization problem (3) is approximated by a sequence of minimizations of linear quadratic cost functions. We define $\mathbf{x}_0^{(k)}$ to be the k^{th} estimate to the solution and linearize the cost function (3) around the model trajectory forecast from this estimate. The next estimate is then defined by

$$\mathbf{x}_0^{(k+1)} = \mathbf{x}_0^{(k)} + \delta\mathbf{x}_0^{(k)}; \tag{6}$$

where the perturbation $\delta\mathbf{x}_0^{(k)} \in \mathbb{R}^n$ is a solution of the linearised cost function

$$\begin{aligned} \mathcal{J}^{(k)}(\delta\mathbf{x}_0^{(k)}) &= \frac{1}{2}(\delta\mathbf{x}_0^{(k)} - [\mathbf{x}^b - \mathbf{x}_0{}^{(k)}])^{\mathrm{T}} \mathbf{B}^{-1}(\delta\mathbf{x}_0^{(k)} - [\mathbf{x}^b - \mathbf{x}_0{}^{(k)}]) \\ &+ \frac{1}{2}\sum_{i=0}^{N} (\mathbf{H}_i \delta\mathbf{x}_i^{(k)} - \mathbf{d}_i^{(k)})^{\mathrm{T}} \mathbf{R}_i^{-1}(\mathbf{H}_i \delta\mathbf{x}_i^{(k)} - \mathbf{d}_i^{(k)}). \end{aligned} \tag{7}$$

Here $\mathbf{d}_i^{(k)} = \mathbf{y}_i - \mathcal{H}_i(\mathbf{x}_i^{(k)})$, where $\mathbf{x}_i^{(k)}$ is the nonlinear trajectory calculated from the current estimate at the initial time using the nonlinear model equation (1). The perturbation $\delta\mathbf{x}_i$ satisfies the linear dynamical equation

$$\delta\mathbf{x}_{i+1} = \mathbf{M}_i \delta\mathbf{x}_i. \tag{8}$$

The linearized observation operator $\mathbf{H}_i$ and the tangent linear model operator $\mathbf{M}_i$ are evaluated at the current estimate of the nonlinear trajectory, usually called the linearization state. The minimization (7) is referred to as the *inner loop*, while the update of the nonlinear model trajectory $\mathbf{x}_i^{(k)}$ is the *outer loop*. On each iteration of the inner loop the TLM is integrated to calculate the evolution of the perturbation, in order to calculate the cost function (7), and the adjoint model is integrated to provide the gradient. A major advantage of the incremental approach is that the inner loop minimization problem may be solved in a smaller dimensional space than the outer loop step, for example at a lower spatial resolution. In this way the TLM and adjoint model need only be run at the lower resolution on each inner loop iteration, while the linearization trajectory from the nonlinear model is still calculated at the higher resolution on each outer loop. The computational savings made by implementing the inner loop in this way made incremental 4D-Var feasible for operational weather and ocean forecasting.

The incremental method was later shown to be equivalent to an inexact Gauss-Newton method applied to the original nonlinear cost function (3) [58]. The outer loop iterations can be shown town tangeinresolution

3 Practical implementation

3.1 Model development

The development of a 4D-Var scheme for the large models used in operational weather and ocean forecasting is often a huge undertaking. In most cases the nonlinear model code already exists and has been developed over many years. These models are very large pieces of software, with maybe close to one million lines of code. In order to develop an incremental 4D-Var scheme the code for the TLM and adjoint model must first be written. The development of a TLM code and adjoint model code from the source code of a nonlinear model is a fairly automatic procedure. The correct code for the TLM can be found from a linearization of each statement of the nonlinear model source code, based on treating the nonlinear model as a series of arithmetic operations and applying the chain rule. The adjoint model is then found by a line-by-line transpose of the TLM source code in reverse order. This method is known as automatic differentiation. We do not go into details of its application here, but refer the reader to several good introductions in the literature [18], [8], [85], [33]. The automatic nature of this procedure has led to many software tools being developed that will produce a TLM and adjoint model code from a nonlinear mode source code. These automatic differentiation tools, or automatic adjoint compilers, are now available commercially for many different programming languages. [2]

In practice the TLM and adjoint models of many large environmental models have been developed by hand, rather than using the automatic compilers. There are several reasons for this. The first is that in many cases of operational weather and ocean forecasting the complexity of the already exisiting nonlinear model codes was such that simple application of the automatic compilers was not possible. In many cases, particularly for large codes developed by many people, it is necessary to tidy the nonlinear model codes to make them suitable for use with the automatic compilers. Many centres felt that the effort to do this would have been greater than coding the TLM and adjoint model by hand.

The second reason for developing the TLM and adjoint codes by hand arises from the nature of the incremental approach to variational data assimilation. Since the TLM and adjoint are run at a lower resolution in the inner loop, the TLM is already an approximate linearization of the nonlinear model used in the outer loop. It is therefore justifiable to make further simplifications in the TLM, in order to reduce the computational cost. As long as the adjoint model is derived from the approximate TLM, then the inner loop minimization will contain the correct gradient information for convergence. In coding the models by hand it is easier to make such simplifications based on physical arguments. For example, many meteorological models contain parametrizations of sub-grid-scale processes (known as the *physics* in the meteorological literature), which include such things as clouds, precipitation and surface drag. The schemes used to represent these processes can be highly complex and often in-

[2] The term *automatic differentiation* refers to the approach itself, not just to the automatic tools.

clude non-di erentiable functions, such as on-o switches. While it is possible for automatic di erentiation to deal with such functions it is usually felt that this level of compexity is not necessary in the TLM and adjoint model. Hence a series of simpler parametrizations have been developed solely for use in incremental 4D-Var, that capture the main behaviour of the more complex schemes [100], [51], [82], [74].

An alternative approach, devised by the Met O ce, is to start from the premise that the linear model must evolve nite and not in nitesimal perturbations and so there is no need for the linear model to be tangent to any nonlinear model. In this approach the linear model is designed with this in mind. In particular, the resolved dynamics is approximated by a discretization of the linearized continuous equations, with various simpli cations in the equations and the discretization. Then simpli ed parametrizations can be used to represent sub-grid-scale processes [72], [60]. The adjoint model is derived from this approximate linear model by the process of automatic di erentiation, ensuring that it provides the exact gradient of the discrete linear cost function.

An essential part of the development of the linear and adjoint models is their testing, as any small mistakes could lead to lack of convergence of the minimization algorithms. Robust tests exist to check the coding of a TLM and adjoint model. The test for the TLM is based on comparing the evolution of a perturbation in the TLM with the evolution of the same perturbation in the nonlinear model. A Taylor series expansion of the nonlinear model operator shows that the evolutions should be closer together as the perturbation size is

this matrix determine the relationships between increments to di erent physical variables or between increments at di erent spatial points. Thus this matrix is fundamental in allowing information to be inferred about unobserved variables or unobserved regions. However, it is usually impossible to represent this matrix in matrix form. If the state vector is of size n then the matrix $\mathbf{B}$ is of size $n \times n$ and when n is of order 10^8 this matrix is impossible to calculate or store. Instead the action of this matrix is usually represented by a variable transform.

We consider the variable transform in the context of incremental variational data assimilation, since that is how it is usually implemented. We de ne a new variable $\mathbf{z}_i \in \mathrm{R}^n$ and a transformation matrix $\mathbf{U}_i \in \mathrm{R}^{n \quad n}$, such that

$$\mathbf{x}_i = \mathbf{U}_i \ \mathbf{z}_i; \qquad i = 0;:::;N: \tag{10}$$

In terms of this new variable the incremental cost function (7) can be written

$$\mathcal{J}^{(k)}(\ \mathbf{z}_0^{(k)}$$

related. For the vertical correlations a transformation to the eigenvectors of a vertical error covariance matrix is used, with the assumption that the errors associated with each eigenvector are uncorrelated. A scaling transformation is also needed to ensure that the variance of the transformed variables is equal to

3.3 Observation errors

As well as representing the errors in the background field it is important to treat properly the errors in the observations within a variational data assimilation system. Observational data received into operational weather and ocean forecasting centres can contain errors from a variety of sources, including limitations in the measuring instrument, biases in the measurements and errors simply due to human error in recording the measurement. The theory of variational data assimilation assumes that all observational errors are random, unbiased errors with a Gaussian distribution and known covariance. It is therefore important that as many of these sources of error as possible are accounted for in the data assimilation system.

A first essential step in an operational data assimilation system is to perform a quality control check on the data themselves. This may consist of several stages. First a check for obvious errors is made, so that if, for example, a ship observation is reported over a land point it will be rejected from the assimilation. Then a so-called `background check' may be made to see how close the observation is to the forecast background field. If the difference from the background is too large when compared with its expected error variance then the observation may be rejected and not used in the assimilation [1]. Once this check has been performed the next step is to identify observations that may have gross errors. This can be done either outside or within the assimilation process. Outside the assimilation each observation can be checked against nearby observations and

data are thinned so that many fewer of them are used [24]. The reasons for this are the difficulty in calculating what the error correlations should be and the difficulty in then representing these correlations within an assimilation scheme in a way that the inverse correlation matrix can easily be applied. To estimate the correlations in satellite data the methods that have mainly been used are a comparison with independent measurements from radiosondes, based on the method of [45], and the use of diagnostics calculated from the data assimilation system itself, based on [26]. Various ways of then representing these correlations within the data assimilation system have been proposed, including the use of a circulant matrix [43], an eigenvalue decomposition [27] and a Markov matrix [88]. However there is so far little use of these methods in operational practice.

3.4 Optimization methods

The minimization of the inner loop cost function (7) requires the use of a suitable optimization algorithm. For the large problems of environmental modelling there are two particularly important constraints. The first is that because of the number of variables in the system it is not possible to obtain second derivative information. The Hessian or second derivative matrix would contain of the order 10^{16} elements, which is impossible to calculate or to store. Hence only methods that require first derivative information can be used. The second constraint is that often these problems must be solved within a real-time forecasting system and hence the computer time that can be used to solve the problem is very limited. Hence the methods much use as few function evaluations as possible. This means that usually the problem is not allowed to run to full convergence and the use of any line search algorithms is prohibitively expensive. Traditionally the algorithms that have most been used within data assimilation systems are quasi-Newton algorithms and conjugate gradient or related Lanczos algorithms. The mathematical details of these algorithms are well explained elsewhere (e.g. [78]) and so here we limit discussion to their implementation in data assimilation systems.

An essential aspect of the minimization procedure for variational data assimilation is an appropriate preconditioning. Experimental evidence indicates that that the Hessian of the inner loop cost function (7) is badly cond(condeT5o)28(w)2[[(similation)-05

O ce [41]. This can be explained by theoretical bounds obtained by [39], [41] that show that the condition number of the transformed problem increases as the spacing between observations decreases and as observations become more accurate. Hence ideally a second level of preconditioning is required after the variable transformation has been performed.

In order to implement a further preconditioning some knowledge of the Hessian (12) of the transformed cost function is required. One way that this can be obtained is by using a Lanczos algorithm to perform the inner loop minimization. The Lanczos method produces estimates of the leading eigenvectors and eigenvalues of the Hessian of the function being minimized. If the rst m eigenvalues $_j$ and eigenvectors $\mathbf{u}_j; j = 1;:::;m$ have su ciently converged then the Hessian (12) can be approximated by the expression

$$\mathbf{I} + \sum_{j=1}^{m} (\ _j - 1)\mathbf{u}_j \mathbf{u}_j^T : \tag{18}$$

This expression can then be used for the preconditioning of subsequent minimizations, under the assumption that the Hessian does not change greatly between one minimization and another [29], [94]. This method, known as spectral preconditioning, is used in the operational forecast system of ECMWF, where three outer loops are performed for each assimilation. During the rst inner loop minimization the Lanczos vectors are stored and these are then used to precondition the mininimization of the second and third inner loop cost functions [29]. It has been shown that this preconditioner belongs to a larger class of limited memory preconditioners [94]. The authors of [94] propose an alternative preconditioner from the same class, based on the Ritz pairs of the Hessian. They found that this can provide an improvement over spectral preconditioning when the estimates of the Hessian eigenpairs are inaccurate. A similar result was also found in the Regional Ocean Modelling System (ROMS), in which both of these preconditioners are implemented [76]. One drawback of both of these methods is that, in order to generate the required information, the rst minimization must be performed before any preconditioning can be applied. So far little attention has been paid to preconditioning of this rst minimization.

With any minimization method it is important to specify appropriate stopping criteria and this is also the case in variational data assimilation. It has been proved that the inner-loop steps of the Gauss-Newton method need to be solved to su cient accuracy in order to ensure convergence of the outer loops [34]. The theory has been used to show how it is natural to use an inner-loop stopping criterion based on the relative change in the norm of the gradient [59]. The tolerance used to stop the iterations must therefore be chosen carefully. If it is too high then there is no guarantee that the outer loop steps will converge. However the convergence should not be pushed below the level of noise on the observations, as then small spatial scales are adjusted to t the observational noise [55]. In many practical forecasting problems such care is not always taken and other criteria are introduced. There are two main reasons for this. One is that in a time-critical forecasting system it may considered more important

and so the background term can be written in the form

$$\mathcal{J}_b(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T \mathbf{B}_w{}^1 \mathbf{w}; \tag{20}$$

where the covariance matrix $\mathbf{B}_w$ is taken to be the diagonal matrix of eigen-

transform to be used in the de nition of the spatial background error covariances described in section 3.2, which then enforces zero boundary increments [69]. However, observational information close to the boundaries can be di cult to use, since the nested model cannot use observations lying outside the domain and the analysis inside the domain may not be consistent with the boundary conditions provided [3], [36]. This can lead to features being arti cially cut-o close to the boundaries.

The alternative approach is to estimate the boundary variables within the assimilation procedure [37], [54], [38]. In this way observations inside the nested domain can update the boundary values and so it is possible to ensure that the analysis is consistent throughout the domain. However in this case it is no longer possible to apply a sine transform to impose the spatial background error covariances. In order to be able to apply a spectral transformation an extension zone is created around the domain to obtain elds that are horizontally periodic. A Fourier transform can then be applied. One di culty in analysing the boundaries in this way is that the lateral boundary conditions are only updated during the assimilation period. During the subsequent forecast no updates are available and the values from the parent model must be used, so there is some inconsistency between the boundary conditions of the analysis and those of the forecast.

The second challenge we consider is the di erence in the spatial scales that can be represented in the nested and parent models. In particular, since the nested model often covers only a small domain, the assimilation scheme is not able to analyse adequately scales of the size of the domain and larger. In applications such as weather prediction it is important to capture these larger scales, since the physical system is inherently multiscale, with strong feedbacks between large and small scales. Hence attempts have been made to improve the large-scale information in nested model data assimilation by providing information on these scales from a parent model analysis. For example, the Met O ce experimented with a system that combined large scale increments from a parent model analysis with the small scale increments from the nested model analysis [3]. In this method the large scales of the nested model analysis are forced to be equal to those of the parent model. An alternative, proposed by [36], is to use the large scales of the parent analysis over the nested model domain as a weak constraint on the variational problem. This is done by adding an extra term to the inner loop cost function (7) that measures the distance between the large scales of the global analysis and those forecast by the nested model. This means that the analysis is constrained by large scales from the parent model, through this additional term, and by large scales from the nested model, through the background term. In theory this should introduce another term including the cross-correlation between these two sources of information. However, in their demonstration of the method in a 3D-Var scheme of the ALADIN model at Meteo-France the authors of [36] concluded that this correlation could be neglected, though at the cost of some inaccuracy.

A more theoretical study of this problem was carried out by [7]. They used a spectral analysis to show how information from waves longer than the domain

size is projected onto different scales in the nested model domain, corresponding to the lowest wave numbers that can be represented on this domain. They demonstrated that by giving more weight to these scales in the background term of the cost function it was possible to retain more of the large scale information from a parent model background. In this method only the large spatial scales from the parent model are used as a constraint in the assimilation, as in [3], but they are not imposed exactly and may be altered by the assimilation process. The authors of [7] demonstrated benefit from this in an idealised system, but the method has not been tested in a realistic model.

3.7 Weak constraint variational assimilation

The formulation of variational data assimilation presented in section 2 assumes that the discrete dynamical model (1) is an exact representation of the physical system being observed. In practice we know that the models contain errors, caused by limitations in our knowledge of the physical equations and limitations in the numerical modelling, such as the need for sub-grid scale parametrizations. In theory it is possible to account for and estimate such errors in variational data assimilation, though implementation in practice is more complicated. We assume an additive error to the model equations, so that the true dynamical system can be written

$$\mathbf{x}_{i+1} = \mathcal{M}_i(\mathbf{x}_i) + \boldsymbol{\eta}_i, \tag{25}$$

where $\boldsymbol{\eta}_i$ is the unknown model error at time t_i. Then we can define a weak constraint 4D-Var problem, in which the model equations do not have to be exactly satisfied over the assimilation window. We define a cost function of the form

$$\begin{aligned} \mathcal{J}(\mathbf{x}_0, \boldsymbol{\eta}_0, \ldots, \boldsymbol{\eta}_{N-1}) &= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^{\mathrm{T}} \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) \\ &+ \frac{1}{2}\sum_{i=0}^{N} (\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i)^{\mathrm{T}} \mathbf{R}_i^{-1}(\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i) + \frac{1}{2}\sum_{i=0}^{N-1} \boldsymbol{\eta}_i^{\mathrm{T}} \mathbf{Q}_i^{-1} \boldsymbol{\eta}_i \end{aligned} \tag{26}$$

subject to (25), where $\mathbf{Q}_i$ is the covariance matrix associated with the model errors $\boldsymbol{\eta}_i$. The weak constraint problem is then to minimize (26) with respect to the initial state $\mathbf{x}_0$ and all the model errors $\boldsymbol{\eta}_i$.

An alternative formulation of the weak constraint problem (26) is to write it in terms of the model state $\mathbf{x}_i$ at each time t_i rather than in terms of the model errors. This leads to the cost function

$$\begin{aligned} \mathcal{J}(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N) &= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^{\mathrm{T}} \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) \\ &+ \frac{1}{2}\sum_{i=0}^{N} (\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i)^{\mathrm{T}} \mathbf{R}_i^{-1}(\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i) \\ &+ \frac{1}{2}\sum_{i=0}^{N-1} (\mathbf{x}_{i+1} - \mathcal{M}_i(\mathbf{x}_i))^{T} \mathbf{Q}_i^{-1}(\mathbf{x}_{i+1} - \mathcal{M}_i(\mathbf{x}_i)), \end{aligned} \tag{27}$$

which is minimized subject to (25). Although (27) is mathematically equivalent

quickly from one analysis cycle to the next.

Despite these initial successes much more work is needed. One particular difficulty is that it is not clear how to differentiate between model bias and observation bias, since the assimilation only measures the difference between the model and the observations. [93] showed a case study of observation bias being interpreted as a model error by weak-constraint 4D-Var. This problem was discussed further by [64] in the context of ocean data assimilation. They suggested that to estimate both model and observation bias it is necessary to include information on the spatial and temporal structure of these biases in the covariance matrices.

In order to then move away from the assumption of a constant bias and treat time-varying systematic and random model errors, more sophisticated methods for describing the evolution of errors must be developed. This evolution is likely to be dependent on the specific model being used, yet general methods for representing this are also needed. At the same time efficient and accurate representations of the covariances of these model errors must be found. The use of the weak-constraint formulation of 4D-Var holds much promise to counteract the inadequacies of models, but many challenges remain open to be able to implement this in very large environmental models

4 Summary and future perspectives

Variational data assimilation is now a well-established method for combining observational data with very large environmental models. However, as has been illustrated in this article, its successful implementation requires careful and judicious choices in each aspect of the assimilation scheme. In some cases these choices are determined by the physical system being modelled or the observational data available, such as the specification of the error covariances in the system. In other cases the choices may be determined by the size of the problem and the need to solve it in an efficient manner, often for real-time forecasting, or by features of the numerical model itself, such as lateral boundary conditions. In each instance the choices to be made will inevitably be a compromise between the ideal solution and what is practically feasible in a given system. We have presented some of the solutions that have been found that have allowed variational data assimilation to be implemented in large environmental forecasting systems. Nevertheless much research continues to improve on these solutions so as to find better estimates of the state and so produce better forecasts.

One particularly active area in numerical weather prediction is the desire to use more information from ensembles of forecasts to provide time-varying covariances for the background errors, combining the advantages of ensemble methods with the advantages of 4D-Var. ECMWF have implemented a system in which an ensemble of 4D-Var assimilations are run and the statistics from this ensemble are used to update the variances of the background errors [10]. Extensions to this method to calculate also the covariance information are being sought. An alternative approach is to use information from ensembles of

forecasts to calculate covariance information throughout the whole assimilation window. This method was proposed by [67] and tested in a global weather prediction model by [14], [15]. An advantage of this method is that the tangent linear and adjoint models are not required in the 4D-Var, since all the evolution information comes through the ensemble of nonlinear model forecasts. Hence this makes development of the system much easier.

Besides the many great challenges that we have discussed in this article, new challenges are arising for the future evolution of variational data assimilation systems. The advent of massively parallel computers means that the algorithms used currently to solve the assimilation problem may no longer be e cient on future computer architectures. Hence work is needed to develop new algorithms to solve the problem, particularly with respect to e cient minimization and preconditioning methods. This may be easier as systems move to a weak-constraint form of 4D-Var but, as discussed above, that introduces its own di culties [30]. Another challenge comes from the move towards more integrated Earth-system models, with di erent environmental models coupled to each other. For example, for seasonal to decadal prediction it is now common to use coupled atmosphere-ocean models, but the initialization of these models with data assimilation is still in its infancy. Particular problems arise from the very di erent time scales in the atmosphere and ocean system and from the model biases in atmosphere and ocean models. Some work has been done to implement 4D-Var in such systems in order to estimate the ocean state and coupling parameters [89], [75], but the estimation of the complete state in coupled atmosphere-ocean models remains an open problem for the coming years.

References

[1] E. Andersson and H. Jarvinen. Variational quality control. *Quart. J. Roy. Meteor. Soc.*, 125(554):697{722, 1999.

[2] T. Auligne, A.P. McNally, and D.P. Dee. Adaptive bias correction for satellite data in a numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, 133(624):631{642, 2007.

[3] S. Ballard, Z. Li, M. Dixon, S. Swarbrick, O. Stiller, and H. Lean. Development of 1-4km resolution data assimilation for nowcasting at the met o ce. In *World Weather Research Program Symposium on Nowcasting and Very Short Range Forecasting (WSN05)*, page Paper 3.02, 2005.

[4] R.N. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. i: Characteristics and measurements of forecast error covariances. *Quart. J. Roy. Meteor. Soc.*, 134(637):1951{1970, 2008.

[5] R.N. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. ii: Modelling the forecast error

covariance statistics. *Quart. J. Roy. Meteor. Soc.*, 134(637):1971{1996, 2008.

[6] R.N. Bannister and M.J.P. Cullen. A regime-dependent balanced control variable based on potential vorticity. In *Proceedings of ECMWF workshop in ow-dependent aspects of data assimilation*, pages 1{13, 2007.

[7] G.M. Baxter, S.L. Dance, A.S. Lawless, and N.K. Nichols. Four-dimensional variational data assimilation for high resolution nested models. *Computers & Fluids*, 46(1):137{141, 2011.

[8] C. Bischof, A. Carle, G. Corliss, A. Griewank, and P. Hovland. ADIFOR: Generating derivative codes from Fortran programs. *Scienti c Programming*, 1:11{29, 1992.

[9] C. Boess, A.S. Lawless, N.K. Nichols, and A. Bunse-Gerstner. State estimation using model order reduction for unstable systems. *Computers & Fluids*, 46(1):155{160, 2011.

[10] M. Bonavita, L. Isaksen, and E. Holm. On the use of eda background error variances in the ecmwf 4d-var. *Quart. J. Roy. Meteor. Soc.*, page doi: 10.1002/qj.1899, 2012.

[11] N. Bormann and P. Bauer. Estimates of spatial and interchannel

[17] Y. Cao, J. Zhu, I.M. Navon, and Z. Luo. A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition. *International Journal for Numerical Methods in Fluids*, 53(10):1571{1583, 2007.

[18] W.C. Chao and L-P. Chang. Development of a four-dimensional variational analysis system using the adjoint method at GLA. Part I: Dynamics. *Mon. Wea. Rev.*, 120:1661{1673, 1992.

[19] P. Courtier, J-N. Thepaut, and A. Hollingsworth. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.*, 120:1367{1387, 1994.

[20] M.J.P. Cullen. Four-dimensional variational data assimilation: A new formulation of the background-error covariance matrix based on a potential-vorticity representation. *Quart. J. Roy. Meteor. Soc.*, 129:2777{2796, 2003.

[21] M.J.P. Cullen. A demonstration of 4d-var using a time-distributed background term. *Quart. J. Roy. Meteor. Soc.*, 136(650):1301{1315, 2010.

[22] D.N. Daescu and I.M. Navon. A dual-weighted approach to order reduc-

[30] M. Fisher, Y. Tremolet, H. Auvinen, D. Tan, and P. Poli. Weak-constraint and long-window 4D-Var. Technical memorandum 655, ECMWF, 2011.

[31] P. Gauthier, C. Charette, L. Fillion, P. Koclas, and S. Laroche. Implementation of a 3d variational data assimilation system at the canadian meteorological centre. part i: The global analysis. *Atmosphere-Ocean*, 37(2):103{156, 1999.

[32] P. Gauthier, M. Tanguay, S. Laroche, S. Pellerin, and J. Morneau. Extension of 3dvar to 4dvar: Implementation of 4dvar at the meteorological service of canada. *Mon. Wea. Rev.*, 135(6):2339{2354, 2007.

[33] R. Giering and T. Kaminski. Recipes for adjoint code construction. *ACM Trans. On Math. Software*, 24:437{474, 1998.

[34] S. Gratton, A.S. Lawless, and N.K. Nichols. Approximate gauss-newton methods for nonlinear least squares problems. *SIAM J. Optim.*, 18(1):106{132, 2007.

[35] A.K. Gri th and N.K Nichols. Adjoint methods in data assimilation for estimating model error. *Flow, Turbulence and Combustion*, 65:469{488, 2000.

[36] V. Guidard and C. Fischer. Introducing the coupling information in a limited-area variational assimilation. *Quart. J. Roy. Meteor. Soc.*, 134(632):723{735, 2008.

[37] N. Gustafsson, L. Berre, S. Hørnquist, X.Y. Huang, M. Lindskog,

[42] B.A. Harris and G. Kelly. A satellite radiance-bias correction scheme for data assimilation. *Quart. J. Roy. Meteor. Soc.*, 127(574):1453{1468, 2001.

[66] M. Lindskog, D. Dee, Y. Tremolet, E. Andersson, G. Radnoti, and M. Fisher. A weak-constraint four-dimensional variational analysis system in the stratosphere. *Quart. J. Roy. Meteor. Soc.*, 135(640):695{706, 2009.

[67] C. Liu, Q. Xiao, and B. Wang. An ensemble-based four-dimensional variational data assimilation scheme. part i: Technical formulation and preliminary test. *Mon. Wea. Rev.*, 136(9):3363{3373, 2008.

[68] A.C. Lorenc. Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, 112:1177{1194, 1986.

[69] A.C. Lorenc. Development of an operational variational assimilation scheme. *J. Met. Soc. Japan*, 75:339{346, 1997.

[79] S.K. Park and D. Zupanski. Four-dimensional variational data assimilation for mesoscale and storm-scale applications. *Meteorology and Atmospheric Physics*, 82(1):173{208, 2003.

[80] D.F. Parrish and J.C. Derber. The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.*, 120:1747{1763, 1992.

[81] F. Rabier and P. Courtier. Four-dimensional assimilation in the presence of baroclinic instability. *Quart. J. Roy. Meteor. Soc.*, 118:649{672, 1992.

[82] F. Rabier, H. Jarvinen, E. Klinker, J.F. Mahfouf, and A. Simmons. The ecmwf operational implementation of four-dimensional variational assimilation. i: Experimental results with simpli ed physics. *Quart. J. Roy. Meteor. Soc.*, 126(564):1143{1170, 2000.

[83] F. Rawlins, S.P. Ballard, K.J. Bovis, A.M. Clayton, D. Li, G.W. Inverarity, A.C. Lorenc, and T.J. Payne. The met o ce global four-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteor. Soc.*, 133(623):347{362, 2007.

[84] C. Robert, S. Durbiano, E. Blayo, J. Verron, J. Blum, and F.X. Le Dimet. A reduced-order strategy for 4d-var data assimilation. *Journal of Marine Systems*, 57(1):70{82, 2005.

[85] N. Rostaing, S. Dalmas, and A. Galligo. Automatic di erentiation in Odyssee. *Tellus*, 45A:558{568, 1993.

[86] Y. Sasaki. An objective analysis based on the variational method. *J. Met. Soc. Japan*, 36:77{88, 1958.

[87] Y. Sasaki. Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev.*, 98:875{883, 1970.

[88] L.M. Stewart. *Correlated observation errors in data assimilation*. PhD thesis, Department of Mathematics, University of Reading, 2010.

[89] N. Sugiura, T. Awaji, S. Masuda, T. Mochizuki, T. Toyoda, T. Miyama, H. Igarashi, and Y. Ishikawa. Development of a four-dimensional variational coupled data assimilation system for enhanced analysis and prediction of seasonal to interannual climate variations. *Journal of Geophysical Research*, 113(C10):C10017, 2008.

[90] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quart. J. Roy. Meteor. Soc.*, 113:1311{1328, 1987.

[91] J-N. Thepaut and P. Courtier. Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Quart. J. Roy. Meteor. Soc.*, 117:1225{1254, 1991.

[92] Y. Tremolet. Accounting for an imperfect model in 4d-var. *Quart. J. Roy. Meteor. Soc.*, 132(621):2483{2504, 2006.

[93] Y. Tremolet. Model-error estimation in 4d-var. *Quart. J. Roy. Meteor. Soc.*, 133(626):1267{1280, 2007.

[94] J. Tshimanga, S. Gratton, A.T. Weaver, and A. Sartenaer. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. *Quart. J. Roy. Meteor. Soc.*, 134(632):751{769, 2008.

[95] J. Vialard, A.T. Weaver, D.L.T. Anderson, and P. Delecluse. Three-and four-dimensional variational assimilation with a general circulation model of the tropical paci c ocean. part ii: Physical validation. *Mon. Wea. Rev.*, 131(7):1379{1395, 2003.

[96] A. Weaver and P. Courtier. Correlation modelling on the sphere using a generalized di usion equation. *Quart. J. Roy. Meteor. Soc.*, 127(575):1815{1846, 2001.

[97] A.T. Weaver, C. Deltel, E. Machu, S. Ricci, and N. Daget. A multivariate balance operator for variational ocean data assimilation. *Quart. J. Roy. Meteor. Soc.*, 131(613):3605{3625, 2005.

[98] A.T. Weaver and I. Mirouze. On the di usion equation and its application to isotropic and anisotropic correlation modelling in variational assimilation. *Quart. J. Roy. Meteor. Soc.*, page doi: 10.1002/qj.1955, 2012.

[99] A.T. Weaver, J. Vialard, and D.L.T. Anderson. Three-and four-dimensional variational assimilation with a general circulation model of the tropical paci c ocean. part i: Formulation, internal diagnostics, and consistency checks. *Mon. Wea. Rev.*, 131(7):1360{1378, 2003.

[100] Q. Xu. Generalized adjoint for physical processes with parameterized discontinuities. Part I: Basic issues and heuristic examples. *J. Atmos. Sci.*, 53:1123{1155, 1996.

[101] D. Zupanski. A general weak constraint applicable to operational 4DVAR data assimilation systems. *Mon. Wea. Rev.*, 125:2274{2292, 1997.