QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **134**: 1971–1996 (2008)
Published online 24 November 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.340

RMetS
Royal Meteorological Society

Review
A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics

R. N. Bannister*
Data Assimilation Research Centre, University of Reading, UK

ABSTRACT: This article reviews a range of leading methods to model the background error covariance matrix (the **B**-matrix) in modern variational data assimilation systems. Owing partly to its very large rank, the **B**-matrix is impossible to use in an explicit fashion in an operational setting and so methods have been sought to model its important properties in a practical way. Because the **B**-matrix is such an important component of a data assimilation system, a large effort has been made in recent years to improve its formulation. Operational variational assimilation systems use a form of control variable transform to model **B**. This transform relates variables that exist in the assimilation's 'control space' to variables in the forecast model's physical space. The mathematical basis on which the control variable transform allows the **B**-matrix to be modelled is reviewed from first principles, and examples of existing transforms are brought together from the literature. The method allows a large rank matrix to be represented by a relatively small number of parameters, and it is shown how information that is not provided explicitly is filled in. Methods use dynamical properties of the atmosphere (e.g. balance relationships) and make assumptions about the way that background errors are spatially correlated (e.g. homogeneity and isotropy in the horizontal). It is also common to assume that the **B**-matrix is static. The way that these, and other, assumptions are built into systems is shown.
The article gives an example of how a current method performs. An important part of this article is a discussion of some new ideas that have been proposed to improve the method. Examples include how a more appropriate use of balance relations can be made, how errors in the moist variables can be treated and how assumptions of homogeneity/isotropy and the otherwise static property of the **B**-matrix can be relaxed. Key developments in the application of dynamics, wavelets, recursive filters and flow-dependent methods are reviewed. The article ends with a round up of the methods and a discussion of future challenges that the field will need to address. Copyright © 2008 Royal Meteorological Society

KEY WORDS balance; control variable transforms; flow dependency; multivariate

Received 12 February 2008; Revised 23 August 2008; Accepted 2 October 2008

1. Introduction to background error covariance modelling

Data assimilation techniques used for meteorological applications, such as forecasting the weather, rely on the availability of error covariance statistics of short numerical forecasts of the atmospheric state. These statistics contain a wealth of information about the nature of meteorological systems and the limitations of the models used to forecast them. They are used in data assimilation in the form of a background (or forecast) error covariance matrix. As is shown in the companion article (Bannister, 2008, hereafter referred to as Part I), the background error covariance matrix describes important characteristics of the probability density function (PDF) of forecast errors, which are usually taken to be normally distributed about the background state. This PDF has a profound impact on the analysed fields: it specifies the importance of the a priori state used, it spreads information from each observation to neighbouring locations and to other variables resulting in a smooth analysis state, it allows sets of observations to complement each other's effect, and it helps the analysis to maintain a state close to balance. This article is a review of some important approaches that have been used or considered in recent years to allow background error covariance matrices to be used effectively in variational data assimilation (VAR) systems, especially in those that are used for numerical weather prediction (NWP).

The 'true' background error covariance matrix for a particular analysis time is denoted $\mathbf{P}^{\mathrm{f}}$. In principle, this matrix may be found by evaluating the Kalman filter equations (Part I), but this calculation is unfeasible for systems used in NWP. This difficulty is largely because of the size of $\mathbf{P}^{\mathrm{f}}$, as NWP systems use a large dimensional state space, typically 10^7 dimensions or more, and so $\mathbf{P}^{\mathrm{f}}$ will have 10^{14} elements, which is

**Correspondence to: R. N. Bannister Department of Meteorology, University of Reading, Earley Gate, Whiteknights, Reading, Berkshire, RG6 6BB, UK.
E-mail: r.n.bannister@reading.ac.uk

prohibitive. In practice, many types of approximation are used to make the problem tractable, which involve replacing $\mathbf{P}^{\mathrm{f}}$ with an approximate form, called $\mathbf{B}$.

- Unlike $\mathbf{P}^{\mathrm{f}}$, $\mathbf{B}$ is usually static and so has no flow dependency. It is assumed though that the static matrix contains some very important properties that $\mathbf{P}^{\mathrm{f}}$ has on average. Rudimentary flow dependency, however, is often included in special ways in VAR, and important examples are discussed in this article. Even though a static matrix needs to be computed only once, it still has the same size as $\mathbf{P}^{\mathrm{f}}$.
- The matrix size problem can be solved partially by modelling the multivariate parts of the covariances. Instead of using explicit covariances between different variables (see Part I for some examples), these covariances can be implied using physical relationships between variables (e.g. balance relationships).
- Further reductions in the amount of explicit covariance information that is needed can be made by modelling the point-to-point spatial covariances. For instance, assumptions are often (but not always) made of homogeneity and isotropy.

Techniques to measure the statistics of $\mathbf{B}$ are described in Part I, but this article is concerned with the methods of constructing approximate but feasible $\mathbf{B}$ matrices, using a process called background error covariance modelling (Fisher, 2003). It is known from data assimilation theory that the closer the $\mathbf{B}$-matrix is to the 'true' forecast error covariance matrix (e.g. $\mathbf{P}^{\mathrm{f}}$ from the Kalman filter if the system is linear), the closer the analysis is to that produced from the optimal combination of background and observational information (e.g. Daley, 1991; Lorenc, 2003a; Fisher, 2007). The ways that the $\mathbf{B}$-matrix can affect the analysed fields to achieve this is covered in Section 3 of Part I. Developments made to the $\mathbf{B}$-matrix used in NWP centres over the years are thought to have contributed significantly to improvements of the accuracy of analyses and forecasts. For example, the superiority of VAR systems (which exploit the techniques described in this article) over their predecessors, for example, optimal interpolation (which do not), is partly due to the use of important dynamical information contained in the $\mathbf{B}$-matrix. This was found, for example,, on the introduction of 3d-VAR at the Met Office (Lorenc *et al.*, 2000). Simmons (2003) reports on forecast improvements of the European Centre for Medium Range Weather Forecasts (ECMWF), some of which coincide with changes made to their $\mathbf{B}$-matrix. It is therefore important to understand methods of background error covariance modelling that are used at present, to allow further improvements to be made. It is the purpose of this article to draw together literature in this subject for these purposes.

Part II has the following structure. In Section 2, we introduce the method of control variable transforms (CVTs) as used in most operational VAR systems. In Sections 3 and 4, we discuss how multivariate and spatial univariate aspects of the problem can be modelled within this framework. In Section 5, we show how the CVTs can be calibrated. In Section 6, we compare the amount of information needed to define CVTs to that in the full $\mathbf{B}$-matrix and give an example covariance structure that is implied by the standard transforms. In Section 7, we review some developments to the basic control variable transform method. The main focuses of Sections 3 and 4 are on the ECMWF and the United Kingdom Met Office systems, but links to other systems are pointed out where appropriate.

2. Modelling B using control variable transforms

The method of CVTs is an effective means of modelling multivariate and univariate aspects of $\mathbf{B}$ approximately in a very compact and efficient way. Use of the method was reported as early as 1992, in the Spectral Statistical Interpolation scheme of the National Meteorological Center, now the National Center for Environmental Prediction (Parrish and Derber, 1992), and has been developed since then. The compactness of $\mathbf{B}$ when it is represented using CVTs is just one motivation of the method. There are also other benefits, which are mentioned at the end of this section. Even though CVTs are based on a mathematical footing, their application to the modelling of error covariances in practice requires physical knowledge, as well as a sense of the phenomenology of error covariances (examples are given in Part I). There are differences in the way that CVTs are implemented between NWP centres.

2.1. Basic method of control variable transforms

2.1.1. Transforming the cost function to new variables

The principle is to make a change of variable that simplifies the background term in the cost function. In the notation of Ide *et al.* (1997), the cost function, J, in the incremental formulation (Courtier, Thépaut and Hollingsworth, 1994) is

$$\begin{aligned} J(\delta\mathbf{x}, \mathbf{x}^{\mathrm{g}}) &= \frac{1}{2}(\delta\mathbf{x} - \delta\mathbf{x}^{\mathrm{b}})^{\mathrm{T}}\mathbf{B}^{-1}(\delta\mathbf{x} - \delta\mathbf{x}^{\mathrm{b}}) \\ &\quad + \frac{1}{2}\left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \delta\mathbf{x})\right\}^{\mathrm{T}} \\ &\quad \times \mathbf{R}^{-1}\left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \delta\mathbf{x})\right\}, \end{aligned} \tag{1}$$

which is minimized with respect to $\delta\mathbf{x}$. Other symbols are as follows: $\mathbf{x}^{\mathrm{g}}$ is a known reference state from which all incremental variables are specified, $\delta\mathbf{x}^{\mathrm{b}}$ is the background increment ($\mathbf{x}^{\mathrm{b}} = \mathbf{x}^{\mathrm{g}} + \delta\mathbf{x}^{\mathrm{b}}$), $\mathbf{y}^{\mathrm{o}}$ is the vector of observations, H is the forward model and $\mathbf{R}$ is the observation error covariance matrix. In this formulation, $\mathbf{x}^{\mathrm{b}}$, $\mathbf{x}^{\mathrm{g}}$, $\delta\mathbf{x}$, $\delta\mathbf{x}^{\mathrm{b}}$ and $\mathbf{B}$ are valid at $t = 0$. Equation (1) is 4d-VAR if H includes a forecast model step to account for observations made at $t > 0$ (e.g. $H \rightarrow H_t M_{t\leftarrow 0}$ for observations at time t where $M_{t\leftarrow 0}$ is the forecast step), as done in Johnson, Hoskins and Nichols (2005), or 3d-VAR otherwise. The analysis, $\mathbf{x}^{\mathrm{a}}$, follows from the $\delta\mathbf{x}$ that minimizes J using

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{g}} + \delta\mathbf{x}, \tag{2}$$

DOI: 10.1002/qj

which is usually used as the initial conditions for a full weather forecast.

The principle is to make a change of variable that simplifies the background term in (1). This change replaces $\delta\mathbf{x}$ with a 'control variable' denoted $\boldsymbol{\chi}$ (implicitly an incremental quantity so the δ-notation is dropped), which is a vector related to $\delta\mathbf{x}$ by $\mathbf{B}^{1/2}$:

$$\delta\mathbf{x} = \mathbf{B}^{1/2}\chi. \tag{3}$$

The operator $\mathbf{B}^{1/2}$ is the CVT and its square-root form allows the background term in the cost function to simplify considerably. Substituting (3) into (1), and noting that $\mathbf{B}$ in (1) can be expressed in terms of its square root as $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{\mathrm{T}/2}$, allows $\mathbf{B}$ to cancel in the background term. The cost function in terms of χ then takes the form

$$\begin{aligned} J(\chi, \mathbf{x}^{\mathrm{g}}) &= \frac{1}{2}(\chi - \chi^{\mathrm{b}})^{\mathrm{T}}(\chi - \chi^{\mathrm{b}}) \\ &\quad + \frac{1}{2}\left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \mathbf{B}^{1/2}\chi)\right\}^{\mathrm{T}} \\ &\quad \times \mathbf{R}^{-1}\left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \mathbf{B}^{1/2}\chi)\right\}, \end{aligned} \tag{4}$$

where χ^{b} is related to $\delta\mathbf{x}^{\mathrm{b}}$ by $\delta\mathbf{x}^{\mathrm{b}} = \mathbf{B}^{1/2}\chi^{\mathrm{b}}$. Note that $\mathbf{B}^{\mathrm{T}/2} = (\mathbf{B}^{1/2})^{\mathrm{T}}$, and even though $\mathbf{B}$ itself is symmetric, $\mathbf{B}^{1/2}$ need not be. In the χ representation (4), the background error covariance matrix becomes the identity matrix and is then trivial to deal with. In other words, components of background error are uncorrelated in the χ representation and have unit variance. It is the transformed cost function (4) that is minimized with respect to χ in VAR. Minimization is achieved using the so-called adjoint method (e.g. Le Dimet and Talagrand, 1986), which requires calculation of the gradient of J with respect to χ. This gradient has the form

$$\begin{aligned} \nabla_{\chi} J &= (\chi - \chi^{\mathrm{b}}) + \mathbf{B}^{\mathrm{T}/2}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \\ &\quad \times \left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \mathbf{B}^{1/2}\chi)\right\}, \end{aligned} \tag{5}$$

where $\nabla_{\chi} J$ is a column vector of derivatives with respect to each component of χ, that is, $\nabla_{\chi} J = (\partial J/\partial\chi_1, \ldots, \partial J/\partial\chi_n)^{\mathrm{T}}$ for n components of χ, and $\mathbf{H}$ is the linearized observation operator introduced in Part I. The state of χ that gives the minimum value of J is used to recover the analysis using (2) and (3), $\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{g}} + \mathbf{B}^{1/2}\chi$. Transformation (3) has removed explicit reference to $\mathbf{B}$ from the background term, but it has not been removed from the problem. In the new problem (4) and (5), background error covariance information is moved to the transform, which is needed to compute $\mathbf{x}^{\mathrm{a}}$, and to compute the observation term at each iteration. The observation operator H acts on model states and not directly on the control variable χ, and also depends upon the reference state $\mathbf{x}^{\mathrm{g}}$, which is not transformed into the control variable representation, as transformations are made only on perturbation quantities such as $\delta\mathbf{x}$.

Vector χ may contain the same number of degrees of freedom as are in $\delta\mathbf{x}$, but this is not necessary. In non-hydrostatic models, for example, meteorologically unimportant acoustic-like modes are represented in $\delta\mathbf{x}$ but are often excluded in χ, and so will not be analysed. This leaves the meteorologically important Rossby and gravity modes for analysis in VAR. Examples of CVTs are shown in Section 3 to demonstrate how this is done. It is also possible to include more elements in χ than are in $\delta\mathbf{x}$, which is permitted as long as the CVT is a valid square root (as revealed by studying, for example, the implied covariance matrix; see Section 2.1.3). Examples of such CVTs are shown in Section 7.

The length of χ may be extended to also include extra information that does not contribute directly to the analysis in the way shown in (2) and (3). Using ideas pursued by Dee (2005) it is possible to include bias-correction parameters of the forward models and/or observations used in the assimilation. Minimizing the cost function with respect to all parts of χ will then yield simultaneously analysis increments (which have been found with bias-corrected models and data) and the bias estimates themselves. This is essentially a data assimilation system with weakly constrained forward models, and is a similar approach to that used in weak-constraint 4d-VAR (e.g. Zupanski, 1997; Zupanski *et al.*, 2005). These extra control variable elements are not discussed further in this article.

The general approach to CVTs is shown schematically in Figure 1. When expressed in terms of model variables (left part of Figure 1) the data assimilation problem is

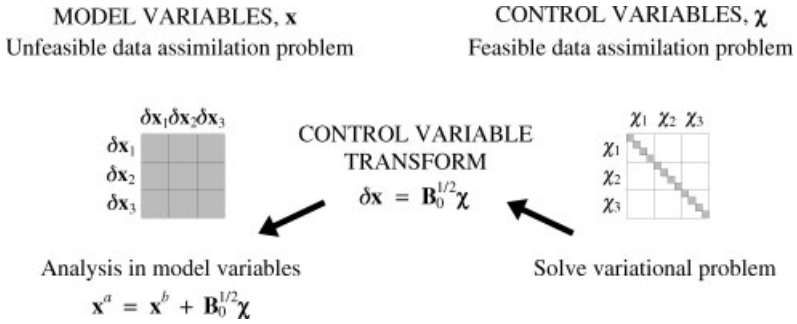

Figure 1. Schematic diagram of the method of control variable transforms (using an example with three model and control variables). The matrices shown are the background error covariance matrices in model variables (left) and control variables (right). Background errors only in the latter case are used (the identity matrix), but background errors in the former case are implied by the control variable transform. Shaded matrix elements denote non-zero elements.

DOI: 10.1002/qj

unfeasible because of the size of the background error covariance matrix. When expressed in terms of control variables (right part of Figure 1) the problem is feasible owing to the trivial nature of background error covariance in χ variables. In this procedure, successive computations are made of the gradient of J, computed with respect to control variables as in (5), and once minimized, the analysis is then expressed in model variables with (2) and (3).

2.1.2. *The control variable transform 'model'*

As $\mathbf{B}$ is not known, $\mathbf{B}^{1/2}$ is also not known, and so (3)–(5) cannot be computed as they stand, but an approximation of $\mathbf{B}^{1/2}$ is constructed to take its place. Let the linear operator $\mathbf{B}_0^{1/2}$ represent an approximation (or model) of $\mathbf{B}^{1/2}$ that we can construct on physical and statistical grounds (this replacement to $\mathbf{B}^{1/2}$ has already been used in Figure 1). The cost function, its gradient and the CVT are then approximated by

$$\begin{aligned} J(\chi, \mathbf{x}^{\mathrm{g}}) &\approx \frac{1}{2}(\chi - \chi^{\mathrm{b}})^{\mathrm{T}}(\chi - \chi^{\mathrm{b}}) \\ &\quad + \frac{1}{2}\left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \mathbf{B}_0^{1/2}\chi)\right\}^{\mathrm{T}} \mathbf{R}^{-1} \\ &\quad \times \left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \mathbf{B}_0^{1/2}\chi)\right\}, \qquad (6) \\ \nabla_\chi J &\approx (\chi - \chi^{\mathrm{b}}) + \mathbf{B}_0^{\mathrm{T}/2}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \\ &\quad \times \left\{\mathbf{y}^{\mathrm{o}} - H(\mathbf{x}^{\mathrm{g}} + \mathbf{B}_0^{1/2}\chi)\right\}, \qquad (7) \\ \delta\mathbf{x} &= \mathbf{B}_0^{1/2}\chi. \qquad (8) \end{aligned}$$

Even though the CVT is now an approximation to $\mathbf{B}^{1/2}$, components of the background error in terms of the control variable associated with $\mathbf{B}_0^{1/2}$ are still assumed to be uncorrellated and to have unit variance. The approximation $\mathbf{B}_0^{1/2}$ is the CVT actually used in VAR. Approximations (6) and (7) will be good if $\mathbf{B}_0^{1/2}$ is a good representation of $\mathbf{B}^{1/2}$, and building such a model is the topic of Sections 3 and 4.

2.1.3. *Implied background error covariance matrix*

It is possible to study $\mathbf{B}_0^{1/2}$ (if this transformation is known) by looking at components of the implied error covariance matrix denoted $\mathbf{B}^{\mathrm{ic}}$. This is the effective background error covariance matrix (in the $\mathbf{x}$-representation) that follows from minimization of (6). The conditions imposed to compute this are the following:

- the background error covariances take the form of the identity in the χ-representation, $\langle \chi\chi^{\mathrm{T}}\rangle = \mathbf{I}$ (where χ is a population of deviations from the 'truth');
- the CVT in (8).

Considering $\delta\mathbf{x}$ in the following to mean a population of deviations from the 'truth' in the $\mathbf{x}$-representation (denoted $\boldsymbol{\eta}$ in Part I), the implied background error covariance matrix is then

$$\mathbf{B}^{\mathrm{ic}} = \langle \delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}}\rangle = \mathbf{B}_0^{1/2}\langle \chi\chi^{\mathrm{T}}\rangle \mathbf{B}_0^{\mathrm{T}/2} = \mathbf{B}_0^{1/2}\mathbf{B}_0^{\mathrm{T}/2}, \qquad (9)$$

where $\langle\rangle$ indicates averaging, and would equal $\mathbf{B}$ if the CVT, $\mathbf{B}_0^{1/2}$, is an exact square root of $\mathbf{B}$. As its name suggests, the implied covariance matrix is not computed explicitly in the assimilation, but its effect on chosen states can be studied by acting with the sequence of operators $\mathbf{B}_0^{1/2}\mathbf{B}_0^{\mathrm{T}/2}$. This is a useful way of looking at the properties of a given CVT model and is used in studies (e.g. Ingleby, 2001; Bannister, 2007).

The challenge of background error covariance modelling is to capture in $\mathbf{B}_0^{1/2}$ the known important features of the background error covariance matrix (see Part I). These features can be studied with the aid of the implied covariance matrix or in assimilation trials. Computational efficiency is important, as $\mathbf{B}_0^{1/2}$ (and its adjoint $\mathbf{B}_0^{\mathrm{T}/2}$) are used in every VAR iteration in (6) and (7).

Different centres have different formulations of $\mathbf{B}_0^{1/2}$, each with its own set of approximations. Table I lists some NWP VAR systems and some of the characteristics of their model of $\mathbf{B}$. These are discussed more fully in the course of this article. The CVT technique is also used in some ocean VAR systems (e.g. Weaver *et al.*, 2005). In Sections 3 and 4, we introduce and compare the ECMWF and the Met Office approaches as examples. Centres use different symbols for $\mathbf{B}_0^{1/2}$: for example, $\mathbf{C}$ has been used by the (previously named) National Meteorological Center, $\mathbf{L}$ is often used at the ECMWF, $\mathbf{U}$ is used at the Met Office, and (confusingly) $\mathbf{U}^{-1}$ is used in the High Resolution Limited Area Model (HIRLAM).

2.2. A generic form of control variable transform

In general, a covariance matrix, $\mathbf{B}$, can be written as an eigenvector decomposition (e.g. Lanczos, 1988)

$$\mathbf{B} = \mathbf{F}\Lambda\mathbf{F}^{\mathrm{T}} = \mathbf{F}\Lambda^{1/2}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\Lambda^{1/2}\mathbf{F}^{\mathrm{T}}. \qquad (10)$$

Here, columns of $\mathbf{F}$ are eigenvectors of $\mathbf{B}$, which are mutually orthogonal ($\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}$), and Λ is the diagonal matrix of eigenvalues, which represent the forecast error variances associated with the eigenvectors. $\mathbf{Q}$ is an arbitrary orthogonal rotation satisfying $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$, and so $\mathbf{Q}$ cancels in (10). It has been added to generalize the expression for the square root (see below). The eigenrepresentation is useful because Λ is diagonal, indicating that the eigenvectors are mutually uncorrelated. Equation (10) allows $\mathbf{B}$ to be decomposed into its square roots. Comparing $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{\mathrm{T}/2}$ with (10) gives an exact form of the CVT

$$\mathbf{B}^{1/2} = \mathbf{F}\Lambda^{1/2}\mathbf{Q}^{\mathrm{T}}. \qquad (11)$$

Such a square-root form helps to ensure that $\mathbf{B}$ is a symmetric and positive semi-definite matrix (Gaspari and Cohn, 1999). There are an infinite number of square-root operators, and each can be represented by a different $\mathbf{Q}$-matrix in (11), which is the reason for including $\mathbf{Q}$ in (10). Possible choices are $\mathbf{Q} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{F}$; however, $\mathbf{Q}$ need not be square, but must satisfy the orthogonality condition. It is not possible, of course, to evaluate

DOI: 10.1002/qj

Table I. Properties of some important VAR systems used in NWP as described in published material. The acronyms used are: European Centre for Medium Range Weather Forecasts (ECMWF), National Meteorological Centre (NMC), National Center for Environmental Prediction (NCEP), World Research and Forecasting system (WRF), National Center for Atmospheric Research (NCAR), Mesoscale Model 5 (MM5), Japan Meteorological Agency (JMA), High Resolution Limited Area Model (HIRLAM), Middle Atmosphere Model (MAM), Canadian Meteorological Centre (CMC) and Regional Atmospheric Modelling Data Assimilation System (RAMDAS). Other symbols are defined in the text (e.g. subscripts b, u, s and t denote balanced, unbalanced, surface and total, and EOF, LBE and NLBE denote empirical orthogonal function, linear balance equation and nonlinear balance equation, respectively).

VAR system	Reference(s)	Control variables	Mass/wind balance	Spatial filter
ECMWF	Courtier *et al.* (1998), Derber and Bouttier (1999)	$\zeta, \eta_u, (T, p_s)_u, q$	Regression	Spectral (horiz.), EOF (vert.)
Met Office	Lorenc *et al.* (2000), Met Office (1995), Ingleby (2001), Rawlins *et al.* (2007)	ψ, χ, p_u, μ	LBE	Spectral (horiz.), EOF (vert.)
NMC/NCEP	Parrish and Derber (1992), Wu Purser and Parrish (2002), de Pondeca *et al.* (2007)	ζ, η, ϕ_u, q or $\psi, \chi_u, (T, p_s)_u, \mu_{pseudo}$	LBE	Spectral (horiz.), EOF (vert.) or Recursive filter
Météo-France (Arpège)	Sadiki and Fischer (2005)	$\zeta, \eta_u, (T, p_s)_u, q$	Regression	Spectral (horiz.), EOF (vert.)
Météo-France (Aladin)	Berre (2000), Fischer *et al.* (2005)	$\zeta, \eta_u, (T, p_s)_u, q_u$	Regression	Spectral (horiz.), EOF (vert.)
WRF	Skamarock *et al.* (2008), Sun *et al.* (2008)	$\psi, \chi_u, (T, p_s)_u, \mu_{pseudo}$	Regression	Spectral (global), Recursive filter (regional)
NCAR (MM5)	Barker *et al.* (2003), Barker *et al.* (2004)	ψ, χ, p_u, q or μ	Hybrid NLBE/ regression	Recursive filter (horiz.) EOF (vert.)
JMA (global)	JMA (2007)	$\zeta, \eta_u, (T, p_s)_u, \log q$	Regression	Spectral (horiz.), EOF (vert.)
JMA (regional)	Honda *et al.* (2005), JMA (2007)	$(\theta, p_s)_b, u_u, v_u, T_u, \mu_{pseudo}$	Regression	Recursive filter (horiz.), EOF (vert.)
HIRLAM	Gustafsson *et al.* (1999, 2001)	$T, u_u, v_u, \log p_s, q$	LBE	Spectral (horiz.), EOF (vert.)
Canadian MAM	Polavarapu *et al.* (2005)	$\psi, \chi_u, (T, p_s)_u, \log q$	Regression	Spectral (horiz.)
CMC	Gauthier *et al.* (1999), Laroche *et al.* (1999)	ψ, χ, ϕ_u, q	LBE	Spectral (horiz.)
RAMDAS	Zupanski *et al.* (2005)	$u, v, w, \Pi, \theta, q_t$	None	Convolutions (horiz./vert.)

the eigenvectors and eigenvalues of the very large **B**-matrix, so (11) cannot be used to represent $\mathbf{B}^{1/2}$ exactly. However, (11) is still useful as it provides a conceptual framework helping to construct a model, $\mathbf{B}_0^{1/2}$. Equation (11) and its inverse (below) are used in Section 4.

The inverse of (11) is

$$\chi = \mathbf{B}^{-1/2}\delta\mathbf{x} = \mathbf{Q}\Lambda^{-1/2}\mathbf{F}^{\mathrm{T}}\delta\mathbf{x}. \tag{12}$$

This has an important interpretation that can sometimes help in the construction of $\mathbf{B}_0^{1/2}$. This inverse transform projects a model increment, $\delta\mathbf{x}$, onto statistically uncorrelated variables (with $\mathbf{F}^{\mathrm{T}}$) and then normalizes the result (with $\Lambda^{-1/2}$) by dividing by the square root of the variances of the eigenvectors. This is followed by the optional rotation, **Q** (as a result of the normalization above, **Q** does not affect the uncorrelated property of the resulting variables). Given the absence of exact knowledge about **F**, Λ and **Q**, a strategy to approximate these is to mimic these steps in $\mathbf{B}_0^{-1/2}$ using physical arguments to give (at least approximately) uncorrelated variables that have unit variance. $\mathbf{B}_0^{1/2}$ follows by inverting these steps. Examples are shown in Sections 3 and 4.

2.3. A pragmatic choice of control variable transform structure

In most VAR systems, CVTs have a number of stages. Commonly, $\mathbf{B}_0^{1/2}$ and $\mathbf{B}_0^{-1/2}$ have the forms

$$\mathbf{B}_0^{1/2} = \mathbf{K}_p\mathbf{B}_s^{1/2}. \tag{13}$$

$$\mathbf{B}_0^{-1/2} = \mathbf{B}_s^{-1/2}\mathbf{K}_p^{-1}. \tag{14}$$

The notation used in (13) and (14) is close to that used by Derber and Bouttier (1999).

DOI: 10.1002/qj

- $\mathbf{K}_\mathrm{p}$ is called the balance operator or the parameter transform. To explain $\mathbf{K}_\mathrm{p}$, it is easier to describe its inverse. The role of $\mathbf{K}_\mathrm{p}^{-1}$ is to take incremental fields in the $\mathbf{x}$-representation (e.g. vorticity, divergence, temperature, etc., on a latitude, longitude and height grid, or whatever variables a given forecast model uses) and to output new parameters on the same grid, which are to be called $\tilde{\chi}_1$, $\tilde{\chi}_2$, $\tilde{\chi}_3$, etc. Parameters are chosen that are thought to be uncorrelated with each other (or nearly so). The parameters can then be treated as univariate (see Section 2.2 of Part I). The types of parameter that are output by $\mathbf{K}_\mathrm{p}^{-1}$ are shown in Table I, and how the parameter transform works is discussed in Section 3.
- $\mathbf{B}_\mathrm{s}^{1/2}$ is called the spatial transform, which is again described with respect to its inverse. Even though $\tilde{\chi}_1$, $\tilde{\chi}_2$, $\tilde{\chi}_3$, etc., are (approximately) uncorrelated with each other, there remain potentially significant covariances between pairs of positions within each parameter's field. The aim of $\mathbf{B}_\mathrm{s}^{-1/2}$ is to project each parameter onto spatial modes, which are mututally uncorrelated, and then to divide by the square root of the variances of each mode. This is akin to application of (12) to each parameter separately. The result is a set of transformed parameters, χ_1, χ_2, χ_3, etc., which make up the control vector χ and whose background errors are uncorrelated (and have unit variance). Because χ_1, χ_2, χ_3, etc., have each been projected onto spatial modes, they are no longer functions of latitude, longitude and height. How the spatial transform works is discussed in Section 4.

Because $\mathbf{B}_\mathrm{s}^{1/2}$ acts on χ, where χ satisfies $\langle \chi \chi^\mathrm{T} \rangle = \mathbf{I}$, $\mathbf{B}_\mathrm{s}^{1/2}$ may be interpreted as a square root of a covariance matrix. $\mathbf{K}_\mathrm{p}$ does not have this property, which is why this operator has not been given a 'square-root' notation. Putting this information together yields the complete transform from (13), which recovers the model variables from the control variables. In the case of three model variables and parameters

$$\begin{aligned}\begin{pmatrix} \delta\mathbf{x}_1 \\ \delta\mathbf{x}_2 \\ \delta\mathbf{x}_3 \end{pmatrix} &= \mathbf{K}_\mathrm{p}\tilde{\chi} = \mathbf{K}_\mathrm{p}\mathbf{B}_\mathrm{s}^{1/2}\chi, \\ &= \mathbf{K}_\mathrm{p} \begin{pmatrix} \mathbf{B}_{\mathrm{s},\tilde{\chi}_1}^{1/2} & 0 & 0 \\ 0 & \mathbf{B}_{\mathrm{s},\tilde{\chi}_2}^{1/2} & 0 \\ 0 & 0 & \mathbf{B}_{\mathrm{s},\tilde{\chi}_3}^{1/2} \end{pmatrix} \begin{pmatrix} \chi_1 \\ \chi_2 \\ \chi_3 \end{pmatrix}. \end{aligned} \tag{15}$$

In $\mathbf{B}_\mathrm{s}^{1/2}$, there is one spatial transform per parameter $\tilde{\chi}_i$, which is denoted by $\mathbf{B}_{s,\tilde{\chi}_i}^{1/2}$.

As well as $\mathbf{B}_0^{1/2}$, the adjoint, $\mathbf{B}_0^{\mathrm{T}/2}$, is needed in (6) and (7). Special techniques exist to construct adjoint operators (Giering and Kaminski, 1998). The inverse transform, apart from being useful to describe the background error covariance modelling strategy as above, is also needed mainly outside the minimization, and in particular for the task of calibrating the transforms (Section 5). There are often other components in an operational CVT, such as simplification operators that modify the resolution of the fields. Such operators are not discussed here as they distract from the essential ideas.

2.4. Other benefits of control variable transforms

The primary motivation of the CVT is to capture the properties of $\mathbf{B}$ without the need for an explicit matrix, but there are also other benefits.

- $\mathbf{B}_0^{1/2}$ introduces balance constraints to the assimilation via $\mathbf{K}_\mathrm{p}$, which gives the VAR system an element of dynamical consistency (Section 3).
- There is a potential numerical advantage to minimizing a cost function of the structure of (6), rather than of (1). The efficiency of the minimization is associated with the eigenvalue structure of the Hessian (the matrix of second derivatives of J), which describes the local shape of the cost function. An illustrative quantity is the condition number, defined as the ratio of the largest to smallest eigenvalues of the Hessian (a large condition number is associated with inefficient minimization). The Hessian with respect to $\delta\mathbf{x}$ in (1) is $\mathbf{B}^{-1} + \mathbf{H}^\mathrm{T}\mathbf{R}^{-1}\mathbf{H}$, and with respect to χ in (6) is $\mathbf{I} + \mathbf{B}_0^{\mathrm{T}/2}\mathbf{H}^\mathrm{T}\mathbf{R}^{-1}\mathbf{H}\mathbf{B}_0^{1/2}$. The former Hessian has a lowest possible eigenvalue of 0 while the latter has a lowest possible eigenvalue of 1. As long as the background term dominates the Hessian eigenvalue spectrum, the latter problem may therefore be expected to have the lower condition number, leading to a more efficient minimization. For this reason, the CVT is sometimes viewed as a means of partially preconditioning the system (Courtier and Talagrand, 1990). Lorenc (1997) and Gauthier *et al.* (1999) indeed find increased convergence rates in minimizing (6) over (1).

3. Parameter transform

VAR systems make a transformation to parameters as part of the CVT (Section 2.3). We show in Sections 3.1 and 3.2 (by reference to the ECMWF and Met Office systems) how these transformations capture implicitly the multivariate covariances of forecast errors.

Because no exact transformation exists to completely decorrelate variables, meteorological parameters $\tilde{\chi}_1$, $\tilde{\chi}_2$, $\tilde{\chi}_3$, etc., are chosen that are thought to be uncorrelated. A principle that is used to choose parameters is the belief that errors in geostrophically balanced parameters (associated with Rossby modes) in the atmosphere are largely decoupled from errors in unbalanced parameters (associated with inertio-gravity modes). There is a hierarchy of balance relations associated with balanced flow (e.g. McIntyre, 2003), but often the simple linear balance relation is used to describe balance. Linear balance is a

DOI: 10.1002/qj

geostrophic-like balance and relates the balanced part of the wind increment, described by the balanced stream function, $\delta\psi_\mathrm{b}$, with the balanced part of the mass increment, described by the balanced pressure, δp_b

$$\nabla_\mathrm{h} \cdot (f\rho_0 \nabla_\mathrm{h} \delta\psi_\mathrm{b}) - \nabla_\mathrm{h}^2 \delta p_\mathrm{b} = 0. \tag{16}$$

Here, f is the Coriolis parameter, ρ_0 is a reference density and ∇_h is the horizontal gradient operator. The fact that mass and wind variables are subject to such a relation sheds light on how multivariate covariances are formed. Balanced increments may be described by either $\delta\psi_\mathrm{b}$ or δp_b (or an associated parameter such as balanced vorticity, $\delta\zeta_\mathrm{b}$), which may be chosen to be one of the new parameters (e.g. $\tilde{\chi}_1$). The remaining components of the flow are unbalanced (e.g. gravity modes described by the unbalanced wind and unbalanced pressure) and may be chosen to be $\tilde{\chi}_2$ and $\tilde{\chi}_3$ (there are two gravity modes for each Rossby mode; Daley, 1991). The justification for the assumption that balanced and unbalanced components are uncorrelated include the following.

- The Rossby and gravity modes are the normal modes of the linearized equations of motion. As such they (and their errors) evolve independently. This can be seen trivially in the linearized shallow-water equations (Daley, 1991). Nonlinearity and forcing will introduce coupling between the modes, but this is assumed to be negligible in the short forecast comprising the background state, although this has not been investigated in detail.
- Phillips (1986) showed that describing forecast errors in terms of the uncorrelated normal modes of the system leads to a good agreement with forecast errors found by empirical means, such as in studies by Hollingsworth and Lönnberg (1986). Phillips considered the geostrophic modes only, showing that the balanced component can be reasonably treated separately from the unbalanced components. More recently, Žagar, Gustafsson and Källén (2004a) used such a normal mode formulation of the **B**-matrix in a tropical model.
- Existing data assimilation systems that exploit this property have been shown to yield physically reasonable error covariances between model variables. Examples are given in the course of this article.

The choice of parameters, and how such concepts can be extended to variables other than mass and wind is covered in the forthcoming sections, which describe example parameter transforms used by two forecast centres (the ECMWF and the Met Office). The background error covariance schemes are constantly being upgraded, but here we describe the schemes according to the most recent work published in the literature at the time of writing. Some recent developments are described in Section 7.

3.1. ECMWF parameter transform

The parameter transform of the ECMWF data assimilation system is similar to that in other systems, such Météo-France, NCEP and the JMA (see Table I). A description of this system (and the Met Office's system in Section 3.2) helps to further explain the idea of the parameter transform. The parameters, $\tilde{\chi}_i$, used in the ECMWF system are vorticity, $\tilde{\chi}_1 = \delta\tilde{\zeta}$, unbalanced divergence, $\tilde{\chi}_2 = \delta\tilde{\eta}_\mathrm{u}$, unbalanced mass, consisting of temperature and surface pressure, $\tilde{\chi}_3 = (\delta\tilde{T}, \delta\tilde{p}_\mathrm{s})_\mathrm{u}$ (regarded as a single parameter) and specific humidity, $\tilde{\chi}_4 = \tilde{q}$ (Derber and Bouttier, 1999). The parameter transform $\mathbf{K}_\mathrm{p}$ allows the model variables in $\delta\mathbf{x}$ (vorticity, $\delta\zeta$, divergence, $\delta\eta$, mass, $(\delta T, \delta p_\mathrm{s})$ and δq) to be recovered from the parameters

$$\begin{pmatrix} \delta\zeta \\ \delta\eta \\ (\delta T, \delta p_\mathrm{s}) \\ \delta q \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 & 0 & 0 \\ \mathcal{MH} & \mathbf{I} & 0 & 0 \\ \mathcal{NH} & \mathcal{P} & \mathbf{I} & 0 \\ 0 & 0 & 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \delta\tilde{\zeta} \\ \delta\tilde{\eta}_\mathrm{u} \\ (\delta\tilde{T}, \delta\tilde{p}_\mathrm{s})_\mathrm{u} \\ \delta\tilde{q} \end{pmatrix}. \tag{17}$$

The matrix operators are explained below. The matrix has a lower triangular form, which is important to the working of the parameter transform, as described below. Parameters are described in the same coordinate system as the model variables, which at ECMWF system, is the spectral representation.

The leading parameter, $\tilde{\chi}_1 = \delta\tilde{\zeta}$, is used to describe the balanced part of the flow. In (17), it is actually identical to the model variable $\delta\zeta$. Under geostrophic adjustment theory, rotational wind is predominantly 'balanced' under certain horizontal and vertical scale regimes (Haltiner and Williams, 1980). Providing that the horizontal scale is smaller than the Rossby radius (but larger than a 'convective scale' where the notion of balance breaks down), this parameter is an appropriate representation of the balanced component of the flow, which is valid for many synoptic-scale situations. A balance relation such as (16) is then valid, where $\delta\tilde{\zeta} = \nabla_\mathrm{h}^2 \delta\psi_\mathrm{b}$.

The second model variable to recover is $\delta\eta$ (second line of (17)). This has balanced and unbalanced components where the balanced component is considered to be the portion of $\delta\eta$ that is correlated with $\delta\tilde{\zeta}$. A 'balanced divergence' increment can be found from $\delta\tilde{\zeta}$ via the regression operator $\mathcal{MH}$ (see below). The remainder is the unbalanced part of $\delta\eta$, $\delta\tilde{\eta}_\mathrm{u}$, and is chosen to be the second parameter, $\tilde{\chi}_2$. In (17), $\mathcal{H}$ is a horizontal regression operator giving the pressure field that is in balance with vorticity ($\mathcal{H}$ is a statistical formulation of the linear balance equation (16)). Other systems that share this approach are shown in Table I. Balanced pressure is then regressed to balanced divergence via $\mathcal{M}$. There are a number of reasons why parameters $\delta\tilde{\zeta}$ and $\delta\tilde{\eta}_\mathrm{u}$ are assumed to be uncorrelated. The first is mentioned in the bullet points at the beginning of Section 3 and the second is in the treatment defining $\delta\tilde{\eta}_\mathrm{u}$ as that part of $\delta\tilde{\eta}$ that is not related to (and hence uncorrelated with) $\delta\tilde{\zeta}$ in the dataset used to define the regression. In the assimilation, however, there is no guarantee that these steps will lead to exact non-correlation between the parameters.

The third model variable to recover is $(\delta\tilde{T}, \delta\tilde{p}_\mathrm{s})$ (third line of (17)). It is found by the same procedure as

DOI: 10.1002/qj

above. It has a balanced part that is found from $\delta\tilde{\zeta}$ via the operator $\mathcal{NH}$. The unbalanced component is partly found from $\delta\tilde{\eta}_\mathrm{u}$ via the operator $\mathcal{P}$ and the remainder is $(\delta\tilde{T}, \delta\tilde{p}_\mathrm{s})_\mathrm{u}$, which is not related to either $\delta\tilde{\zeta}$ or $\delta\tilde{\eta}_\mathrm{u}$.

The fourth model variable is δq. In (17), this specific humidity increment is taken to be decoupled from the dynamical variables, even though in the real world it is not. The transform could be extended by attempting to relate part of the specific humidity increment to $\delta\tilde{\zeta}$, $\delta\tilde{\eta}_\mathrm{u}$ and $(\delta\tilde{T}, \delta\tilde{p}_\mathrm{s})_\mathrm{u}$, and choosing the residual as the moisture parameter $\tilde{\chi}_4$. The Météo-France Aladin limited area data assimilation system (Berre, 2000) is very similar to that of the ECMWF, but uses such a decoupled specific humidity parameter (Table I). Their version of the fourth line in (17) reads

$$\delta q = \mathcal{Q}\mathcal{H}\delta\tilde{\zeta} + \mathcal{R}\delta\tilde{\eta}_\mathrm{u} + \mathcal{S}(\delta\tilde{T}, \delta\tilde{p}_\mathrm{s})_\mathrm{u} + \delta\tilde{q}_\mathrm{u}. \tag{18}$$

allowing them to be compared to the directly measured covariances (see Part I).

3.2. Met Office parameter transform

The Met Office VAR scheme has a different design to the ECMWF scheme (see Table I). Its transform is

$$\begin{pmatrix} \delta\zeta \\ \delta\eta \\ \delta p \\ \delta T \\ \delta q \end{pmatrix} = \begin{pmatrix} \nabla_\mathrm{h}^2 & 0 & 0 & 0 \\ 0 & \nabla_\mathrm{h}^2 & 0 & 0 \\ \mathcal{H} & 0 & \mathbf{I} & 0 \\ \mathcal{TH} & 0 & \mathcal{T} & 0 \\ (\Gamma + \mathcal{YT})\mathcal{H} & 0 & \Gamma + \mathcal{YT} & \Lambda \end{pmatrix} \begin{pmatrix} \delta\tilde{\psi} \\ \delta\tilde{\chi} \\ \delta\tilde{p}_\mathrm{u} \\ \delta\tilde{\mu} \end{pmatrix}. \tag{21}$$

The variables on the left-hand side are not quite the model variables used by the Met Office forecast model (winds δu and δv are used instead of $\delta\zeta$ and $\delta\eta$), but presentational complications are avoided when $\delta\zeta$ and $\delta\eta$ are used (the Helmholtz relation relates these to winds). This transform is not square, but it is still valid (the number of parameters is fewer than the number of model variables, but the transform can diagnose the missing information).

As (17), (21) is also a lower triangular form, but there are a number of fundamental differences between the ECMWF and Met Office transforms: (1) the choice of parameters are stream function, $\delta\tilde{\psi}$, velocity potential, $\delta\tilde{\chi}$, unbalanced pressure, $\delta\tilde{p}_\mathrm{u}$ and relative humidity, $\delta\tilde{\mu}$; (2) the fields are in a spatial (instead of a spectral) representation; (3) the operators in the matrix are analytical instead of statistical regressions (see Table I for other systems that share this property). Note that $\delta\tilde{\chi}$ is the velocity potential parameter while $\delta\chi$ is the general control parameter. The documentation referenced in Table I does not describe the Met Office's parameter transform in matrix form and so some symbols in (21) have been invented for this article.

In the basic Met Office scheme, the first (balanced) parameter, $\tilde{\chi}_1 = \delta\tilde{\psi}$, gives the vorticity increment by the first line of (21). The second model variable, $\delta\eta$ (second line of (21)), is taken to be wholly unbalanced as it depends on the second parameter, $\delta\tilde{\chi}$, but not on the first, $\delta\tilde{\psi}$. This ignores the existence of a balanced divergence, which is accounted for in the ECMWF. This simplification made at the Met Office may be justified by Obukhov (1954), who showed that the rotational and divergent winds are uncorrelated if the flow is homogeneous.

$$\mathbf{B}^\mathrm{ic} = (\mathbf{K}_\mathrm{p}\mathbf{B}_\mathrm{s}^{1/2})(\mathbf{K}_\mathrm{p}\mathbf{B}_\mathrm{s}^{1/2})^\mathrm{T}, \tag{19}$$

$$= \begin{pmatrix} \mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}} & \mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}}(\mathcal{MH})^\mathrm{T} & \mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}}(\mathcal{NH})^\mathrm{T} & 0 \\ \mathcal{MH}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}} & \mathcal{MH}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}}(\mathcal{MH})^\mathrm{T} + \mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\eta}_\mathrm{u}} & \mathcal{MH}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}}(\mathcal{NH})^\mathrm{T} + \mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\eta}_\mathrm{u}}\mathcal{P}^\mathrm{T} & 0 \\ \mathcal{NH}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}} & \mathcal{NH}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}}(\mathcal{MH})^\mathrm{T} + \mathcal{P}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\eta}_\mathrm{u}} & \mathcal{NH}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\zeta}}(\mathcal{NH})^\mathrm{T} + \mathcal{P}\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\eta}_\mathrm{u}}\mathcal{P}^\mathrm{T} + \mathbf{B}^\mathrm{ic}_{\mathrm{s},(\tilde{T},\tilde{p}_\mathrm{s})_\mathrm{u}} & 0 \\ 0 & 0 & 0 & \mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{q}} \end{pmatrix}. \tag{20}$$

In (18), $\mathcal{Q}$, $\mathcal{R}$ and $\mathcal{S}$ are vertical regression operators, determined by a calibration procedure, and $\delta\tilde{q}_\mathrm{u}$ is the choice for the next control parameter, $\tilde{\chi}_4$.

This procedure can be continued for other degrees of freedom. In systems that assume hydrostatic balance, only three dynamical parameters are important (one balanced and two unbalanced), but this may be extended for high-resolution systems where the hydrostatic balance approximation breaks down. It could also be applied to chemical species such as ozone (Lahoz, Fonteyn and Swinbank, 2007).

The implied covariance matrix between errors in the model variables can be derived for the ECMWF transform using (9), (13) and (17). This yields (19) and (20) (Derber and Bouttier, 1999), where the rows and columns correspond to errors of the model variables $\delta\zeta$, $\delta\eta$, $(\delta T, \delta p_\mathrm{s})$, δq and $\mathbf{B}^\mathrm{ic}_{\mathrm{s},\tilde{\chi}_i} = \mathbf{B}^{1/2}_{\mathrm{s},\tilde{\chi}_i}\mathbf{B}^{\mathrm{T}/2}_{\mathrm{s},\tilde{\chi}_i}$ (see Section 4). The contributions to each covariance term now become evident. For instance, it is clear that the multivariate covariance between $(\delta T, \delta p_\mathrm{s})$ and $\delta\eta$ errors has two parts, $\mathcal{MH}\mathbf{B}_{\mathrm{s},\tilde{\zeta}}(\mathcal{NH})^\mathrm{T}$ and $\mathbf{B}_{\mathrm{s},\tilde{\eta}_\mathrm{u}}\mathcal{P}^\mathrm{T}$. This shows that the ECMWF covariance model couples errors in these variables via the vorticity and unbalanced divergence parameters. The coupling via vorticity may be regarded as the contribution that is due to balanced processes. It is clear that (17) implies no coupling between specific humidity and other variables (Berre, 2000, shows the implied covariances when specific humidity coupling in the form of (18) is included in the CVT). The structures of implied covariances are usually studied numerically,

DOI: 10.1002/qj

The third model variable, δp (third line of (21)), has balanced and unbalanced components. The balanced part is $\mathcal{H}\delta\tilde{\psi}$ and is the solution of the linear balance equation (16), $\mathcal{H}\delta\tilde{\psi} = \nabla_{\mathrm{h}}^{-2}[\nabla_{\mathrm{h}} \cdot (f\rho_0\nabla_{\mathrm{h}}\delta\tilde{\psi})]$, where ∇_{h}^{-2} is the level-by-level Poisson solver. The remaining unbalanced part is the third parameter, $\delta\tilde{p}_{\mathrm{u}}$. Temperature increments (fourth line of (21)) can also be found from the three parameters introduced so far, by assuming hydrostatic balance, which is a good assumption on synoptic scales. Hydrostatic balance allows the temperature increment, δT, to be determined from δp

$$\begin{aligned}\delta T &= \mathcal{T}\delta p \\ &= \frac{g}{c_p}\left(\frac{\partial\Pi_0}{\partial z}\right)^{-1}\left\{\Pi_0\left(\frac{\partial\Pi_0}{\partial z}\right)^{-1}\frac{\partial}{\partial z} - 1\right\}\frac{\kappa\Pi_0}{p_0}\delta p,\end{aligned} \tag{22}$$

where Π_0 and p_0 are reference state Exner pressure and pressure, and other symbols have their usual meanings. In (21), $\mathcal{T}$ is a representation of this differential operator.

The fifth model variable, δq (fifth line of (21)), is found from other variables using thermodynamic arguments. The main moisture parameter is the relative humidity, $\delta\tilde{\mu}$, rather than specific humidity. This is an important difference from the ECWMF's scheme. Given the definition of relative humidity in this scheme, $\mu = q/q_{\mathrm{sat}}$ (where q_{sat} is the saturated specific humidity, which is a function of temperature and pressure), it is possible to linearize it and write specific humidity increments in terms of pressure and temperature increments:

$$\begin{aligned}\delta q &= q_{0\mathrm{sat}}\delta\tilde{\mu} + \mu_0\frac{\partial q_{0\mathrm{sat}}}{\partial p}\delta p + \mu_0\frac{\partial q_{0\mathrm{sat}}}{\partial T}\delta T, \\ &= \Lambda\delta\tilde{\mu} + \Gamma\delta p + \mathcal{Y}\delta T.\end{aligned} \tag{23}$$

The assumed absence of covariances between parameters has reduced the amount of statistical information needed to describe **B**, but the autocovariances of each parameter between different positions in space still need to be dealt with. There are usually in excess of 10^6 components of a field and so the spatial error covariances cannot be dealt with explicitly. The spatial transform is the next stage of the covariance model; see (13)–(15).

From (15), parameters, generically denoted by $\tilde{\chi}_i$, are related to the control variables, χ_i, via the spatial transforms (shown here for three parameters)

$$\begin{pmatrix}\tilde{\chi}_1 \\ \tilde{\chi}_2 \\ \tilde{\chi}_3\end{pmatrix} = \begin{pmatrix}\mathbf{B}_{\mathrm{s},\tilde{\chi}_1}^{1/2} & 0 & 0 \\ 0 & \mathbf{B}_{\mathrm{s},\tilde{\chi}_2}^{1/2} & 0 \\ 0 & 0 & \mathbf{B}_{\mathrm{s},\tilde{\chi}_3}^{1/2}\end{pmatrix}\begin{pmatrix}\chi_1 \\ \chi_2 \\ \chi_3\end{pmatrix}, \tag{25}$$

4. Spatial transforms

The moisture parameter raises non-trivial issues in background error covariance modelling.

The implied covariance matrix (9) can be derived for the Met Office transform using (9), (13) and (24). This yields (24), where the definitions $\mathcal{D} = \Gamma + \mathcal{Y}\mathcal{T}$ and $\mathcal{G} = \mathcal{H}\mathbf{B}_{\tilde{\psi}}^{\mathrm{ic}}\mathcal{H}^{\mathrm{T}} + \mathbf{B}_{\tilde{p}_{\mathrm{u}}}^{\mathrm{ic}}$ are made for shorthand. The rows and columns of (24) correspond to the model variables $\delta\psi$, $\delta\eta$, δp, δT and δq. An example of the Met Office's implied covariances is shown in Section 6.

The consequence of adopting relative, rather than specific humidity, as a parameter implies couplings between errors of the model variables vorticity, pressure and specific humidity in the complicated way shown in (24). This means that in the Met Office's system, perturbations in the specific humidity field (from specific humidity observations) will affect winds, pressure and temperature via the **B**-matrix. This is physically reasonable in conditions that are close to saturation. In the troposphere, for example, this scheme preserves relative humidity (preserving cloud where it is present) when there are no moisture observations, even if VAR changes the temperature. It does this by making compensating changes to the specific humidity in (23) (via the balanced and unbalanced control parameters) consistent with $\delta\tilde{\mu} = 0$. It can otherwise, however, lead to anomalous correlations between variables. In the stratosphere, for example, where the air has a very low relative humidity, there is no physical reason for this coupling to occur (Lahoz and Geer, 2003).

$$\mathbf{B}^{\mathrm{ic}} = \begin{pmatrix}\nabla_{\mathrm{h}}^2\mathbf{B}_{\mathrm{s},\tilde{\psi}}^{\mathrm{ic}}\nabla_{\mathrm{h}}^2 & 0 & \nabla_{\mathrm{h}}^2\mathbf{B}_{\mathrm{s},\tilde{\psi}}^{\mathrm{ic}}\mathcal{H}^{\mathrm{T}} & \nabla_{\mathrm{h}}^2\mathbf{B}_{\tilde{\psi}}^{\mathrm{ic}}\mathcal{H}^{\mathrm{T}}\mathcal{T}^{\mathrm{T}} & \nabla_{\mathrm{h}}^2\mathbf{B}_{\tilde{\psi}}^{\mathrm{ic}}\mathcal{H}^{\mathrm{T}}\mathcal{D}^{\mathrm{T}} \\ 0 & \nabla_{\mathrm{h}}^2\mathbf{B}_{\tilde{\chi}}^{\mathrm{ic}}\nabla_{\mathrm{h}}^2 & 0 & 0 & 0 \\ \mathcal{H}\mathbf{B}_{\tilde{\psi}}^{\mathrm{ic}}\nabla_{\mathrm{h}}^2 & 0 & \mathcal{G} & \mathcal{G}\mathcal{T} & \mathcal{G}\mathcal{D}^{\mathrm{T}} \\ \mathcal{T}\mathcal{H}\mathbf{B}_{\tilde{\psi}}^{\mathrm{ic}}\nabla_{\mathrm{h}}^2 & 0 & \mathcal{T}\mathcal{G} & \mathcal{T}\mathcal{G}\mathcal{T}^{\mathrm{T}} & \mathcal{T}\mathcal{G}\mathcal{D}^{\mathrm{T}} \\ \mathcal{D}\mathcal{H}\mathbf{B}_{\tilde{\psi}}^{\mathrm{ic}}\nabla_{\mathrm{h}}^2 & 0 & \mathcal{D}\mathcal{G} & \mathcal{D}\mathcal{G}\mathcal{T}^{\mathrm{T}} & \mathcal{D}\mathcal{G}\mathcal{D}^{\mathrm{T}} + \Lambda\mathbf{B}_{\tilde{\mu}}^{\mathrm{ic}}\Lambda^{\mathrm{T}}\end{pmatrix}. \tag{24}$$

Here, as usual, subscript '0' refers to reference state quantities. In the actual scheme, there are other terms related to the different molecular weights of air and water, but these are neglected here for simplicity. The matrices Λ, Γ and $\mathcal{Y}$ are shorthand for the operators that appear in the first line of (23). Equation (23) with (22), and lines 3 and 4 of (21), explain the last line of (21).

Studies yield opposing outcomes regarding the validity of relative humidity as a control parameter. Lorenc, Roulstone and White (2003) state that relative humidity errors are more weakly correlated to temperature errors than are specific humidity errors, but Dee and da Silva (2003) report the opposite. These findings indicate that

DOI: 10.1002/qj

and the inverse transform follows. There are a variety of means of modelling spatial covariances, but we start by describing the way that they have been modelled in the ECMWF and Met Office systems.

4.1. ECMWF spatial transform

The sequence of operators that represents the standard ECMWF spatial transform has the following forward and inverse transforms (Derber and Bouttier, 1999)

$$\tilde{\chi}_i = \mathbf{B}_{\mathrm{s},\tilde{\chi}_i}^{1/2}\chi_i = (\mathbf{S}^{-\mathrm{T}}\mathbf{V}^{1/2}\mathbf{S}^{\mathrm{T}})(\mathbf{E}\mathbf{D}^{1/2})\chi_i, \quad (26)$$

$$\chi_i = \mathbf{B}_{\mathrm{s},\tilde{\chi}_i}^{-1/2}\tilde{\chi}_i = (\mathbf{D}^{-1/2}\mathbf{E}^{\mathrm{T}})(\mathbf{S}^{-\mathrm{T}}\mathbf{V}^{-1/2}\mathbf{S}^{\mathrm{T}})\tilde{\chi}_i, \quad (27)$$

where the symbols are defined below. There is one such transform for each parameter, but the $\tilde{\chi}_i$ subscript (which should be present on the component operators in (26) and (27)) is dropped for simplicity of notation. At the ECMWF, each $\tilde{\chi}_i$ is an incremental field of parameter i as a function of horizontal wave number and model level, and χ_i is the same parameter but expressed in a new representation whose background errors are uncorrelated and have unit variance (Figure 1). The spatial transform consists of two parts described here in terms of its inverse (27).

The first part of the inverse transform is the right-hand bracketed part of (27) and accounts for the point-by-point variances of each parameter. Operator $\mathbf{S}^{\mathrm{T}}$ is a level-by-level inverse Fourier transform to the grid representation. This is followed by the diagonal matrix $\mathbf{V}^{-1/2}$ containing the inverse standard deviations of the grid-point parameters. The final operator is the Fourier transform $\mathbf{S}^{-\mathrm{T}}$, which returns to the spectral representation. This transform normalizes the parameters so that they have unit variance at each grid position.

The second part of the inverse transform is the left-hand bracketed part of (27) and deals with the vertical component of error covariances for each wave number. It may be regarded as an application of the generic transform (12) with $\mathbf{Q}$, $\Lambda^{-1/2}$ and $\mathbf{F}^{\mathrm{T}}$ in (12) akin to $\mathbf{I}$, $\mathbf{D}^{-1/2}$ and $\mathbf{E}^{\mathrm{T}}$ in (27), respectively. The rows of $\mathbf{E}^{\mathrm{T}}$ comprise special vertical modes, which may be specified separately for each spectral mode (in practice only for each total wave number). The vertical modes are eigenvectors of vertical error covariance matrices (for each spectral mode) prepared beforehand in a one-off procedure and are sometimes called empirical orthogonal functions (EOFs). Other schemes that share this method of treating vertical covariances are shown in Table I. After projecting onto the vertical modes, the diagonal matrix $\mathbf{D}^{-1/2}$ normalizes the vertical modes so that they have unit variance. The diagonal elements of $\mathbf{D}$ are the eigenvalues of the vertical error covariance matrices mentioned above.

These transforms achieve approximately the aims of the CVT: the non-correlation of errors property is achieved by assuming that errors between spectral modes (and errors between the vertical modes) are uncorrelated, and the unit variance property is achieved by the normalization matrices. There are important consequences, in particular, of making the assumption that spectral modes are uncorrelated (see Section 4.1.2). The matrices $\mathbf{D}$, $\mathbf{E}$ and $\mathbf{V}$ are found via a calibration procedure (Section 5) in which only the variances, $\mathbf{V}$, are allowed to have a seasonal variation. Adding seasonal dependence in this way is a poor substitute for proper flow dependence, but is necessary to capture the seasonality that errors have on average. The $\mathbf{V}$-matrix for parameters $\delta\tilde{\zeta}$ and $\delta\tilde{q}$ is however treated in a different way to other parameters. Instead of being prescribed seasonally, vorticity variances are found from a cycling algorithm, which estimates the propagation of errors from the Hessian of the previous assimilation cycle (Fisher and Courtier, 1995). Specific humidity variances are prescribed as a function of temperature and specific humidity of the background state (Derber and Bouttier, 1999).

There are a number of important features about this ECMWF spatial transform.

- The diagonal matrix $\mathbf{D}^{1/2}$ in (26) acts on states that are a function of spectral mode in the horizontal and vertical modes, and so its diagonal elements comprise a function in that representation. For a given parameter and vertical mode, the diagonal elements of $\mathbf{D}^{1/2}$ plotted as a function of total wave number give a spectrum that determines the length-scale of the error correlations for that parameter and vertical mode. In this spectral covariance model, it is therefore not possible to prescribe via $\mathbf{D}^{1/2}$ horizontal length-scales as a function of model level or horizontal position.
- The matrix $\mathbf{E}$ in (26) transforms the vertical mode representation to model level representation separately for each wave number ($\mathbf{E}^{\mathrm{T}}$ in (27) performs the inverse). There is allowed to be a different set of vertical modes for each wave number. This allows vertical correlations to be dependent upon horizontal scale, but it is then not possible to prescribe vertical correlations as a function of horizontal position.
- The diagonal operator $\mathbf{V}^{1/2}$ in (26) acts in grid space. This ensures that the standard deviation field is well represented in grid space. By analogy with the first bullet point, this means that the variation of standard deviations with wave number or vertical mode cannot be prescribed.
- If the spectral dependence of the vertical modes in $\mathbf{E}$ is weak and the vertical mode dependence of $\mathbf{D}^{1/2}$ is also weak, then the covariance matrix implied by (26) will give rise to 'nearly' separable correlation functions (see Section 4.1.3). In Section 6.3 of Part I, we show that separability is often incompatible with the notion of balance.

Some consequences of this ECMWF transform are not desirable, and in Section 4.3 we review alternatives that try to overcome them. In particular, a wavelet-based scheme (Fisher, 2003, 2004) is now operational in the

DOI: 10.1002/qj

ECMWF system, which allows vertical correlations to acquire a prescribed position as well as scale dependence.

4.1.1. *Implied spatial error covariance matrix of the ECMWF spatial transform*

The spatial part of the ECMWF implied covariance matrix for one parameter is, from (9) and (26),

$$\mathbf{B}_{\mathrm{s}}^{\mathrm{ic}} = \mathbf{B}_{\mathrm{s}}^{1/2}\mathbf{B}_{\mathrm{s}}^{\mathrm{T}/2} = \mathbf{S}^{-\mathrm{T}}\left\{\mathbf{V}^{1/2}\left[\mathbf{S}^{\mathrm{T}}\mathbf{E}\mathbf{D}\mathbf{E}^{\mathrm{T}}\mathbf{S}\right]\mathbf{V}^{1/2}\right\}\mathbf{S}^{-1}, \tag{28}$$

which is an error covariance matrix in the spectral representation. Inside the curly brackets is the implied covariance matrix in the grid representation. This part can be compared to the general form of a covariance matrix (9) of Part I ($\mathbf{B} = \Sigma\mathbf{C}\Sigma$, where Σ is a diagonal matrix of standard deviations and $\mathbf{C}$ is a correlation matrix). This comparison shows that the part of (28) enclosed in square brackets can be interpreted as the implied interspatial correlation matrix.

4.1.2. *Homogeneity and isotropy of horizontal correlations in the ECMWF spatial transform*

The implied covariance matrix (28) is now examined in terms of its homogeneity, isotropy and separability.

For simplicity these properties are examined in a continuous two-dimensional space of horizontal position (r) and height (z) of size L and H, respectively (the spectral transform is then a level-by-level one-dimensional Fourier transform in the horizontal). Let k and ν represent wave number and vertical mode with maximum values K and N, respectively. Then, $E(k, \nu, z)$ is the matrix element of $\mathbf{E}$ corresponding to the weight of vertical mode ν at height z for wave number k, and $D(k, \nu)$ is the diagonal element of $\mathbf{D}$ corresponding to mode (k, ν). The correlation part of (28) (square brackets), acting on a function $x(r, z)$, to give $x'(r', z')$, is expanded as follows:

$$\begin{aligned} x'(r', z') &= \frac{1}{\sqrt{L}}\int_{k=0}^{K} \mathrm{d}k \exp(\mathrm{i}kr') \\ &\quad \times \int_{\nu=0}^{N} \mathrm{d}\nu E(k, \nu, z')D(k, \nu) \\ &\quad \times \int_{z=0}^{H} \mathrm{d}z E(k, \nu, z) \\ &\quad \times \frac{1}{\sqrt{L}}\int_{r=0}^{L} \mathrm{d}r \exp(-\mathrm{i}kr)x(r, z), \\ &= \int \mathrm{d}z \int \mathrm{d}r \left[\frac{1}{L}\int \mathrm{d}k \int \mathrm{d}\nu \exp\{\mathrm{i}k(r' - r)\}\right. \\ &\quad \left. \times E(k, \nu, z')D(k, \nu)E(k, \nu, z)\right]x(r, z). \end{aligned} \tag{29}$$

In the last line, integration limits have been omitted for simplicity. The term in large brackets is the implied correlation matrix element, denoted $C^{\mathrm{ic}}(r', z';\ r, z)$, which is the error correlation of the parameter between positions (r, z) and (r', z'). Importantly, the correlations depend only on separation, $r - r'$, and not individually on r and r'. This is the property of homogeneity that correlations in the ECMWF system have. If $\mathbf{D}$ were not a diagonal matrix, then correlations would not be homogeneous.

Another property that can be analysed in a similar way is isotropy, but this is a property of two dimensions in the horizontal. Correlations between two points in space are isotropic if they are independent of the orientation of the line joining the points. The diagonal elements of $\mathbf{D}$ (which are in spectral space) have a particularly simple form for errors that are isotropic in grid space, where elements of $\mathbf{D}$ are a function of total wave number only, and not of the individual wave number components (Bartello and Mitchell, 1992; Berre, 2000). On a Cartesian grid with wave number components k_x and k_y, the total wave number is $\sqrt{(k_x^2 + k_y^2)}$, and on the globe, the total wave number is the label n of the spherical harmonic Y_n^m.

4.1.3. *Separability of correlations in the ECMWF spatial transform*

Equation (29) is now used to examine separability. Consider the possibility that $E(k, \nu, z)$ is independent of k and $D(k, \nu)$ is independent of ν. Elements of the implied correlation matrix in (29) are then

$$\begin{aligned} C_{\mathrm{s}}^{\mathrm{ic}}(r', z'; r, z) &= \frac{1}{L}\int \mathrm{d}k \int \mathrm{d}\nu \exp\{\mathrm{i}k(r' - r)\} \\ &\quad \times E(\nu, z')D(k)E(\nu, z), \\ &= \frac{1}{L}\left[\int \mathrm{d}k \exp\{\mathrm{i}k(r' - r)\}D(k)\right] \\ &\quad \times \left\{\int \mathrm{d}\nu E(\nu, z')E(\nu, z)\right\}. \end{aligned} \tag{30}$$

By fixing (r, z), $C_{\mathrm{s}}^{\mathrm{ic}}(r', z'; r, z)$ gives correlations as a function of (r', z'). Equation (30) shows that the correlation function can be written as the product of two terms: one a function of r' and not z' (first large bracket), and the other a function of z' and not r'. This property is called separability, and the above shows that the ECMWF transform is capable of being separable under certain conditions. Separability is not always desirable as it can hinder the ability of the covariance model to represent balance (see Part I, Section 6.3).

The properties outlined in Sections 4.1.2 and 4.1.3 apply to parameters, and not to model variables (unless a parameter is also a model variable). The correlation functions for model variables are found via the parameter transform, which also mixes spatial covariances from different parameters. The implied covariance matrix for the complete ECMWF transform has been given as (20).

4.2. Met Office spatial transform

The spatial transform used in the Met Office VAR system (Lorenc *et al.*, 2000) has a similar design to the ECMWF, but there are notable differences. The Met Office analogues of (26) and (27) are as follows for the

DOI: 10.1002/qj

forward and inverse transforms (for consistency here we use a notation close to the ECMWF where appropriate):

$$\tilde{\chi}_i = \mathbf{B}_{\mathrm{s},\tilde{\chi}_i}^{1/2}\chi_i = (\mathbf{E}\mathbf{M}^{1/2})(\mathbf{S}^{\mathrm{T}}\Lambda_{\mathrm{h}}^{1/2})\chi_i, \quad (31)$$

$$\chi_i = \mathbf{B}_{\mathrm{s},\tilde{\chi}_i}^{-1/2}\tilde{\chi}_i = (\Lambda_{\mathrm{h}}^{-1/2}\mathbf{S}^{-\mathrm{T}})(\mathbf{M}^{-1/2}\mathbf{E}^{\mathrm{T}})\tilde{\chi}_i. \quad (32)$$

At the Met Office, $\tilde{\chi}_i$ is an incremental field of parameter i as a function of longitude, latitude and model level and χ_i is the same parameter but expressed in a new representation whose background errors are uncorrelated and have unit variance (Figure 1). As has been done in Section 4.1, this transform is described in terms of the inverse (32). Of the two parts to (32), one is designed to remove vertical correlations and the other is designed to remove horizontal correlations, and each can be understood with reference to the generic transform (12).

The first part is the right-hand bracketed part in (32) and deals with the vertical component of error covariances (the 'vertical transform'). It has a similar form to (12) where $\mathbf{Q}$, $\Lambda^{-1/2}$ and $\mathbf{F}^{\mathrm{T}}$ in (12) akin to $\mathbf{I}$, $\mathbf{M}^{-1/2}$ and $\mathbf{E}^{\mathrm{T}}$ in (32), respectively (a derivation of (31) and (32) is given in Bannister, 2004). The rows of $\mathbf{E}^{\mathrm{T}}$ are the vertical modes (or vertical EOFs). In principle, these may be specified as a function of horizontal position, but the modes used are taken to be constant over the globe, and so are derived from globally averaged statistics (this is done to avoid the problems discussed in Bannister, 2004). After projecting onto the vertical modes, the diagonal matrix $\mathbf{M}^{-1/2}$ acts as a normalization. Matrix $\mathbf{M}$ is allowed to be position-dependent, but in practice it is a function of latitude band only. After the vertical transform, fields are a function of longitude, latitude and vertical mode.

The second part is the left-hand bracketed part in (32) and deals with the horizontal component of error covariances for each vertical mode (the 'horizontal transform'). Terms $\mathbf{Q}$, $\Lambda^{-1/2}$ and $\mathbf{F}^{\mathrm{T}}$ in (12) are akin to $\mathbf{I}$, $\Lambda_{\mathrm{h}}^{-1/2}$ and $\mathbf{S}^{-\mathrm{T}}$ in (32), respectively. Operator $\mathbf{S}^{-\mathrm{T}}$ is a vertical mode-by-vertical mode Fourier transform, and the resulting spectral field is multiplied by elements of the diagonal matrix $\Lambda_{\mathrm{h}}^{-1/2}$. The result is the control parameter, χ_i, in a spectral and vertical mode representation.

The matrices $\mathbf{M}^{1/2}$, $\mathbf{E}$ and $\Lambda_{\mathrm{h}}^{1/2}$ are precomputed via a calibration procedure (Section 5), with some built-in seasonal dependence. Seasonality is simulated by interpolating the statistics between those computed for two times of the year, six months apart (a month or so in winter and in summer). There are similarities and important differences between the designs of the Met Office and ECMWF spatial transforms.

- The operators in the vertical transform depend upon horizontal position (actually just latitude band in $\mathbf{M}$). This means that vertical correlations are prescribed as a function of horizontal position. This is different from the ECMWF transform, which prescribes vertical correlations as a function of scale.
- The diagonal matrix Λ_{h} of the horizontal transform contains weights as a function of spectral and vertical mode. As such, this diagonal matrix performs a similar role to $\mathbf{D}$ in the ECMWF transform.
- There is no explicit grid-point-by-grid-point standard deviation term in the Met Office transform. Instead, these standard deviations are implied by the other terms. Consequently, the separation of implied covariances into correlations and standard deviations, as in (9) of Part I, is not straightforward.
- If the position dependence in $\mathbf{M}^{1/2}$ is weak and the vertical mode dependence of $\Lambda_{\mathrm{h}}^{1/2}$ is also weak, then the covariances implied by (31) will be 'nearly' separable (see Section 4.2.3), as at the ECMWF.

The ability of the Met Office transform to (at least partially) model vertical correlations as a function of position is important with regard to the assimilation of nadir satellite radiances. The coupling between the deep vertical weighting functions of the radiance operator (those rows of the operator $\mathbf{H}$ in (11) of Part I corresponding to satellite radiance observations; e.g. Lerner, Weisz and Kirchengast, 2002) and the vertical structure functions of $\mathbf{B}$ can be important in VAR to use such observational information in an optimal way (see Section 3.4 in Part I). This effect is exploited most fully if the latitude dependence of the vertical covariances are captured by the error covariance models. An example of a marked latitude dependence of vertical correlation structures is Figure 10(b) of Part I (showing temperature error correlations).

4.2.1. *Implied spatial error covariance matrix of the Met Office spatial transform*

The spatial part of the Met Office implied covariance matrix for one parameter, from (9) and (31), is

$$\mathbf{B}_{\mathrm{s}}^{\mathrm{ic}} = \mathbf{B}_{\mathrm{s}}^{1/2}\mathbf{B}_{\mathrm{s}}^{\mathrm{T}/2} = \mathbf{E}\mathbf{M}^{1/2}\mathbf{S}^{\mathrm{T}}\Lambda_{\mathrm{h}}\mathbf{S}\mathbf{M}^{\mathrm{T}/2}\mathbf{E}^{\mathrm{T}}. \quad (33)$$

4.2.2. *Homogeneity of horizontal correlations in the Met Office spatial transform*

Unlike the ECMWF implied spatial covariance matrix, (28), here it is not straightforward to look at only the correlation part. Performing a similar analysis to that in Section 4.1.2, but for covariances (instead of just correlations) gives for the Met Office

$$B_{\mathrm{s}}^{\mathrm{ic}}(r', z'; r, z) = \frac{1}{L}\int \mathrm{d}k \int \mathrm{d}\nu \exp\{\mathrm{i}k(r' - r)\} \times E(\nu, z')M^{1/2}(r', \nu)\Lambda_{\mathrm{h}}(k, \nu)M^{1/2}(r, \nu)E(\nu, z). \quad (34)$$

Here, r is horizontal position, z is height, k is wave number, ν is vertical mode, $B_{\mathrm{s}}^{\mathrm{ic}}(r', z'; r, z)$ is the implied error covariance between positions (r, z) and (r', z'), $E(\nu, z)$ is the weight of the νth vertical mode at height z, $M(r, \nu)$ is the diagonal element of $\mathbf{M}$ corresponding to (r, ν),

DOI: 10.1002/qj

$\Lambda_h(k, \nu)$ is the diagonal element of Λ_h corresponding to (k, ν) and the integrals in (34) are over the complete wave number and vertical mode space (as in Section 4.1.2). The structure functions in (34) are homogeneous only if the r dependence of M is removed. The correlation part of (34) may be derived using (9) of Part I

$$C_s^{ic}(r', z'; r, z) = \frac{B_s^{ic}(r', z'; r, z)}{\sqrt{B_s^{ic}(r, z; r, z) B_s^{ic}(r', z'; r', z')}}, \quad (35)$$

but it is not straightforward to investigate this analytically for properties of homogeneity.

4.2.3. Separability of covariances in the Met Office spatial transform

Equation (34) is now used to examine separability. Consider the possibility that $M(r, \nu)$ is independent of r and $\Lambda_h(k, \nu)$ is independent of ν. Then (34) is

$$B_s^{ic}(r', z'; r, z) = \frac{1}{L}\left[\int \mathrm{d}k \exp\{\mathrm{i}k(r' - r)\}\Lambda_h(k)\right] \times \left\{\int \mathrm{d}\nu E(\nu, z') M(\nu) E(\nu, z)\right\}. \quad (36)$$

In a similar way to the ECMWF correlation functions, the Met Office structure functions are separable under these circumstances.

5. Calibrating the transforms with training data

The spatial transforms $\mathbf{B}_s^{1/2}$ (e.g. (26) or (31) in Section 4) rely on matrices that have to be determined from training data. This process is sometimes called the calibration step. It is often a relatively expensive process and so is often performed only once for a particular system and/or season using a population of forecast error estimates, $\delta\mathbf{x}$. Methods used to determine populations of $\delta\mathbf{x}$ are discussed in Section 5 of Part I.

An example means of determining the component matrices of $\mathbf{B}_s^{1/2}$ and $\mathbf{B}_s^{-1/2}$ is outlined here from a given population of $\delta\mathbf{x}$. We use the simple spatial transforms of Derber and Bouttier (1999) in (26) and (27), but the process can be extended for more complicated forms. The procedure comprises three steps, which make progressive use of the inverse transforms.

- Step 1: determine parameters. Act on each member, $\delta\mathbf{x}$, with $\mathbf{K}_p^{-1}$, to give parameters, $\tilde{\chi}$. The triangular structure of $\mathbf{K}_p$ (e.g. in (17)) means that $\mathbf{K}_p^{-1}$ follows from $\mathbf{K}_p$ in a straightforward manner. The matrices $\mathbf{V}$, $\mathbf{E}$ and $\mathbf{D}$ in (26) and (27) are as yet unknown. There is a set of such matrices for each parameter, but the set for one parameter only, $\tilde{\chi}_i$, is discussed below and steps 2 and 3 must be repeated for the other parameters. (There should be a $\tilde{\chi}_i$ label on each matrix, but this is omitted in the following for ease of notation.) The population mean, $\langle\tilde{\chi}_i\rangle$, may be subtracted from $\tilde{\chi}_i$ if required.
- Step 2: determine $\mathbf{V}$. Act on each member of $\tilde{\chi}_i$ with $\mathbf{S}^T$, which transforms each spectral space parameter to grid space. This is the first operator on the right of (27). Repeating for each population member allows the diagonal matrix of variances, $\mathbf{V}$, to be calculated for each parameter, $\mathbf{V} = \mathrm{diag}\langle(\mathbf{S}^T\tilde{\chi}_i)(\mathbf{S}^T\tilde{\chi}_i)^T\rangle$.
- Step 3: determine $\mathbf{E}$ and $\mathbf{D}$. Act on each member of $\tilde{\chi}_i$ with $(\mathbf{S}^{-T}\mathbf{V}^{-1/2}\mathbf{S}^T)$ (as on the right of (27)). Let $c(\mathbf{k})$ be a vertical column of this field for one wave number $\mathbf{k}$. The vertical error covariance matrix of each column is $\langle c(\mathbf{k})c(\mathbf{k})^T\rangle$. Its eigenvalues form the diagonal components of the part of $\mathbf{D}$ that is associated with $\mathbf{k}$, and eigenvectors form the columns of the part of $\mathbf{E}$ that is associated with $\mathbf{k}$.

This completes the calibration step. In practice, however, some manual adjustment of the matrices may be required to improve the assimilation (e.g. Isaksen, Fisher and Berner, 2007).

6. Efficiency and example implied covariances of the control variable transform method

The modelling of the multivariate and univariate parts of the **B**-matrix with a CVT, as shown in the previous sections, has removed the need to deal with **B** as an explicit matrix. Based on the characteristics published on the ECMWF and Met Office CVTs, Table II gives estimates of the amount of information needed to define their CVTs, and an arithmetic operation count required to act with them. This is compared to the information and operation count for the full **B**-matrix. This table highlights the benefits of the CVT method that have allowed the VAR method to work.

Even though parts of the CVT are derived from forecast errors (as demonstrated in Section 5), the implied **B**-matrix, $\mathbf{B}^{ic}$, as seen by VAR, is only an approximation to **B** due to the simplifications made in the parameter and spatial transforms (Sections 3 and 4). It is straightforward to test $\mathbf{B}^{ic}$ by comparing important correlation structures computed from $\mathbf{B}^{ic}$ against those calculated directly from the forecast errors (see Section 6 of Part I).

This is an important task, but such comparisons are rarely shown. Ingleby (2001) makes such a comparison. As an example, Figure 2 shows vertical temperature correlations (with the ~ 500 hPa level) implied from the Met Office's CVT as a function of latitude. There is no spatial covariance matrix for δT in the Met Office's scheme, but they can be implied from those of $\delta\tilde{\psi}$ and $\delta\tilde{p}_u$ via the expression $\mathcal{T}(\mathcal{H}\mathbf{B}_{\tilde{\psi}}^{ic}\mathcal{H}^T + \mathbf{B}_{\tilde{p}_u}^{ic})\mathcal{T}^T$ (fourth column and row of (24)). The explicit correlations are presented as Figure 10(b) of Part I, also from Ingleby (2001). The underlying structure is reproduced by $\mathbf{B}^{ic}$ (i.e. narrower vertical length-scales in the Tropics than at midlatitudes, and negative correlations with the upper troposphere and lower stratosphere) but the features are too weak. In particular, moving from the Tropics to midlatitudes, $\mathbf{B}^{ic}$ fails to reproduce the dramatic increase

DOI: 10.1002/qj

Table II. Estimate of the number of floating point numbers needed to define the ECMWF- and Met Office-like CVTs (Sections 3 and 4), and the number of arithmetic operations to use them. The operation count has been multiplied by 2 to account for the use of the CVT and its adjoint. Estimates are shown for typical sizes of global systems used in the assimilation and the storage is compared to that of storing all elements of the explicit **B**-matrix.

VAR system	Approx. storage (floating point numbers)				Approx. operation count	
	$\mathbf{K}_{\mathrm{p}}$	$\mathbf{B}_{\mathrm{s}}^{1/2}$	Total	Explicit **B**	$\mathbf{K}_{\mathrm{p}}$	$\mathbf{B}_{\mathrm{s}}^{1/2}$
ECMWF T255 L91	2×10^9	41×10^6	2×10^9	560×10^{12}	4×10^9	39×10^9
Met Office N108 L50	0	65×10^3	65×10^3	50×10^{12}	62×10^6	2×10^9

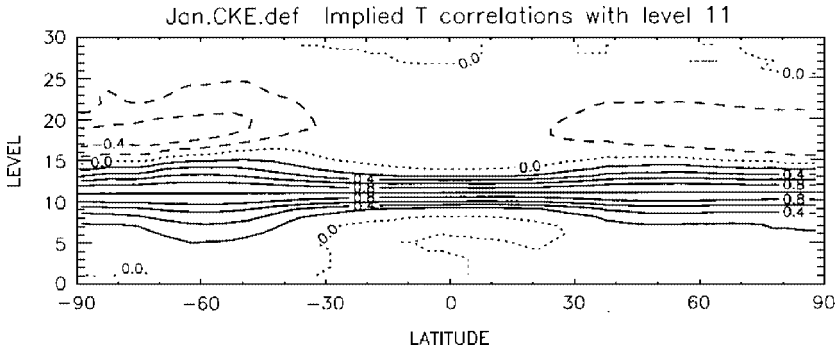

Figure 2. Implied vertical correlations of temperature at ~ 500 hPa with other levels as a function of latitude. ©Crown copyright 2001, from Ingleby (2001). Data supplied by the Met Office.

of vertical length-scale present in **B**. This problem is probably because of the use of globally averaged (rather than latitudinally dependent) vertical modes in the Met Office's representation of $\mathbf{B}_{\tilde{\psi}}^{\mathrm{ic}}$ and $\mathbf{B}_{\tilde{p}_{\mathrm{u}}}^{\mathrm{ic}}$ (Section 4.2). It is also possible to study multivariate aspects of $\mathbf{B}^{\mathrm{ic}}$ found from off-diagonal blocks of (24).

7. Developments in control variable transforms

As explained in general terms in Part I, the **B**-matrix has a striking effect on the analysis increments, so it is important to use a CVT that is based on sound physical principles and gives rise to realistic implied error covariances. The schemes that have been reviewed in Sections 3 and 4 to model multivariate and spatial aspects of covariances are efficient to use (Table II), but have limitations. Here, a number of developments are described, either at the ECMWF, the Met Office or elsewhere. The motivations for improving the CVTs are often based on the following arguments.

- The schemes reviewed in Sections 3 and 4 do not have adequate flow dependence. In the light of the importance of **B**, flow dependence of forecast errors is an important aspect to capture (Kalnay *et al.*, 1997), but remains difficult to model in VAR. Even in 4d-VAR, where the **B**-matrix is propagated to the appropriate times of the observations (Thépaut, Hoffman and Courtier, 1993), the **B**-matrix is reset to its static value at the start of each assimilation cycle (usually every 6 or 12 h). It is possible to capture some flow dependence via revised parameter transforms (Section 7.1), spatial transforms (Section 7.2) and also in other ways (Section 7.3). As an illustration of how the structure of forecast errors evolve, Ménard and Chang (2000) used a Kalman filter to assimilate methane concentrations. They show that forecast errors become contorted by the Kalman filter equations (see (3) in Part I), and analogous results may be expected for meteorological variables. In the absence of a realistic flow-dependent **B**-matrix, essentially the 'wrong' matrix is used in VAR for the particular flow regime, which will lead to a suboptimal analysis.
- Basic schemes generally yield horizontal structure functions of parameters that are based on considerations of homogeneity and isotropy, but the figures in Sections 3 and 6 of Part I, and in other works (e.g. Zhang, 2005), show that error correlations and covariances of many meteorological variables are often inhomogeneous and anisotropic. Fortunately though, the choice of control parameters made in Section 3 reduces the need to model anisotropy in the control parameters using the spatial transforms, as anisotropy may be implied by the parameter transform for some model variables. For instance, the structure functions for rotational and divergent wind parameters (e.g. $\delta\psi$ and $\delta\chi$) may be assumed to be isotropic, but those for zonal and meridional winds implied via the Helmholtz relations ($\delta u = -\partial\delta\psi/\partial y + \partial\delta\chi/\partial x$ and $\delta v = \partial\delta\psi/\partial x + \partial\delta\chi/\partial y$) are highly anisotropic under geostrophic balance (e.g. Kalnay, 2003).
- The ECMWF and Met Office spatial transforms differ in a way that highlights another shortcoming

DOI: 10.1002/qj

Table III. Summary of the key developments of CVTs showing some important advantages and disadvantages.

Development	Section	Advantages	Disadvantages
Flow-dependent balance relations	7.1.1	Nonlinear balance equation gives a more accurate description of balance in regions where the accelerations are high (e.g. strong cyclones).	Not known whether the flow-dependence is similar to that present in $\mathbf{P}^{\mathrm{f}}$.
Potential vorticity-based control variable	7.1.2	Identifies the correct regime to apply mass/wind balance. Allows existence of an unbalanced rotational wind	Requires accurate calibration via solution of three-dimensional elliptic equations. Sensitive to vertical grid staggering.
Balanced vertical motion and divergence	7.1.3	Accounts for a 'balanced' divergence component. Paves the way for calculation of the 'balanced' moisture component.	Requires calibration via solution of three-dimensional elliptic equations.
Treatment of moisture	7.1.4	Accounts for correlations between moisture and other variables. Incorporates knowledge of ascent/descent of air (see above).	Requires reliable diagnosis of vertical motion.
Distorted grids	7.2.1	Relaxes the homogeneous/isotropic assumptions. The geostrophic coordinate transform (GCT) spreads information along fronts, and cyclonic (anticyclonic) regions are contracted (expanded).	Difficult to determine a distortion operator, $\mathbf{L}_\tau$, without the GCT. The GCT is not valid in the Tropics. Implementation is more favourable on isentropic surfaces than on model levels (Lorenc, 2007).
Wavelet formulation	7.2.2	Models covariances as position- and scale-dependent. Allows inhomogeneity and anisotropy, and for different parameter transforms to be used at different scales.	The control vector becomes longer than in the standard CVT. The 'uncertainty principle' causes a trade-off between the possible resolution of position and scale dependency.
Recursive filters	7.2.3	Allows inhomogeneity and anisotropy.	Pole problems can arise due to non-uniform grid. Recursive filters computed along grid directions can give rise to artificial anisotropies.
Diffusion operators	7.2.4	Allows inhomogeneity, anisotropy and complicated boundary conditions.	Becomes less efficient for large length-scales and fine grids.
Reduced rank Kalman filter	7.3.1	Builds in a formal flow-dependent component to an otherwise conventional VAR system.	Requires a Hessian singular vector calculation step. The blending of flow-dependent and static covariances can lead to non-physical structure functions.
Errors of the day	7.3.2	Gives rise to flow-dependent structure functions because of the use of modes bred from the nonlinear model.	The control vector becomes longer than in the standard CVT.

of such simple schemes. In the ECMWF system, vertical correlations are modelled as a function of wave number, yet in the Met Office system, they are modelled as a function of latitude. There is a need to prescribe vertical correlations as a function of wave number and horizontal position simultaneously. This achieves vertical error covariances that are scale- and position-dependent.

Having identified these key points that need to be addressed, some schemes alternative or complementary to those already described in Sections 3 and 4 are reviewed here. Some of the methods mentioned below also exploit the physics of the system, but others are designed to reflect the phenomenological properties of forecast errors. A summary of developments is given in Table III, which lists their respective advantages and disadvantages.

7.1. Developments to the parameter transforms

7.1.1. Flow-dependent balance relations

In the parameter transform (Section 3), the operator $\mathcal{H}$ is used to relate rotational wind to balanced mass, either in an analytic or regressional fashion. In either case, $\mathcal{H}$ is a near-static operator. $\mathcal{H}$ may be replaced with an analytical balance operator with greater sophistication than the linear balance equation. A contender is the nonlinear balance equation (37), but linearized about a reference state taken to be the background (38) (as the CVT schemes require linearity)

$$\nabla_{\mathrm{h}} \cdot \rho_0 \left\{ f \nabla_{\mathrm{h}} \psi + (\mathbf{v}^{\psi} \cdot \nabla_{\mathrm{h}}) \mathbf{v}^{\psi} \right\} - \nabla_{\mathrm{h}}^2 \delta p_{\mathrm{b}} = 0, \quad (37)$$

$$\nabla_{\mathrm{h}} \cdot \rho_0 \left\{ f \nabla_{\mathrm{h}} \delta \psi + (\mathbf{v}_0^{\psi} \cdot \nabla_{\mathrm{h}}) \delta \mathbf{v}^{\psi} + (\delta \mathbf{v}^{\psi} \cdot \nabla_{\mathrm{h}}) \mathbf{v}_0^{\psi} \right\} - \nabla_{\mathrm{h}}^2 \delta p_{\mathrm{b}} = 0. \quad (38)$$

The terms in (38) that do not appear in (16) are cyclostrophic terms and become important when the

DOI: 10.1002/qj

curvature of the background flow is high. These are formulated in terms of the rotational part of the horizontal flow denoted $\mathbf{v}^{\psi}$. These terms change from cycle to cycle, and so including them is a straightforward way of modelling flow-dependent error covariances within the CVT. This has been done by Fisher (2003) and Barker *et al.* (2004) (see Table III, row 1).

7.1.2. A potential vorticity-based balanced control parameter

The use of rotational wind as the leading balanced control parameter (as either $\delta\tilde{\zeta}$ or $\delta\tilde{\psi}$) is questionable. Geostrophic adjustment theory states that rotational wind is 'balanced' only on horizontal scales that are short compared to the Rossby radius, L_{R}. Balanced flow is better described by a mass parameter on horizontal scales that are long compared to L_{R} (as done in the regional HIRLAM and JMA systems; see Table I), and is described by a combination of wind and mass in intermediate regimes. Consider the linearized potential vorticity (PV) perturbation, δQ, for the shallow water equations on an f-plane

$$\delta Q \approx gD\nabla_{\mathrm{h}}^2\delta\psi - f_0\delta\phi, \tag{39}$$

given $\delta\psi$ (wind) and $\delta\phi$ (mass) perturbations, where f_0 is the Coriolis parameter, g is the acceleration due to gravity and D is the vertical length-scale. The PV is relevant because it is associated with the balanced component of the perturbations. On small horizontal and large vertical scales, $\delta Q \sim gD\nabla_{\mathrm{h}}^2\delta\psi$, meaning that PV (and hence balanced motion) is described by rotational wind, and on large horizontal and small vertical scales, $\delta Q \sim -f_0\delta\phi$, meaning that PV is described by mass (e.g. Wlasak, Nichols and Roulstone, 2006). The definition of 'large' and 'small' horizontal scales is made with respect to $L_{\mathrm{R}} = \sqrt{gD}/f_0$ (e.g. Kalnay, 2003). This effect is clear in results by Žagar, Gustafsson and Källén (2004b) who found that in the Tropics, where horizontal length-scales are generally short compared to L_{R}, observations of wind (the balanced variable in this regime) can reconstruct the mass field, but not vice versa.

Cullen (2003) considered this problem in the ECMWF system by introducing a modified parameter, $\tilde{\chi}_1 = \delta\tilde{\zeta}_{\mathrm{b}}$ (the balanced component of vorticity), which describes the PV irrespective of the flow regime. It is equivalent to the existing parameter $\delta\tilde{\zeta}$ (total vorticity) when the horizontal scale is small and when the vertical scale is large. The modified parameter allows for the presence of an unbalanced vorticity component, which is not allowed for in the standard scheme (Section 3.1). The scheme would be expected to perform better than the standard approach, as the balanced and unbalanced parameters are defined more rigorously, allowing the hypothesis of non-correlation between the balanced and unbalanced modes to be exploited (Bannister *et al.*, 2008). Cullen (2003) reports that the scheme works well, but experienced problems with vertical operators due to the Lorenzian grid staggering of the ECMWF model. Bannister and Cullen (2006) are implementing the scheme in the Met Office system, which uses a different (Charney–Phillips) vertical staggering, and Katz (2007) has studied the approach in a simplified shallow water system (see Table III, row 2).

7.1.3. Balanced vertical motion and divergence

As shown in (24), the Met Office's scheme (21) neglects correlations between rotational and divergent components of the horizontal wind, even though in reality part of the divergent wind increment is 'balanced', and so should be expected to be correlated with the rotational wind. This is taken into account at the ECMWF and other centres by the regression operator $\mathcal{M}$ in (17). Another strategy is to use a diagnostic analytical relation between rotational wind and balanced divergence increments. Developments underway attempt this by computing a balanced vertical wind from the rotational wind with the quasi-geostrophic omega equation, and by computing a balanced divergent wind from the vertical wind with the continuity equation (M. Sharpe, personal communication). The new control parameter replacing $\delta\tilde{\chi}$ in (21) would become the residual 'unbalanced' velocity potential increment, $\delta\tilde{\chi}_{\mathrm{u}}$. As with the nonlinear balance equation in Section 7.1.1, the omega equation is also nonlinear and its linearization about the background state will introduce a degree of flow dependence to this part of the **B**-matrix. The use of an analytical omega equation has also been investigated in a similar context by Fisher (2003) and by Pagé, Fillion and Zwack (2007) in high-resolution models. Knowledge of vertical wind is also important in the treatment of moisture (Section 7.1.4; see Table III, row 3).

7.1.4. Advances in the treatment of moisture

Representing how errors in moisture relate to errors in other variables is difficult to model using CVTs. In reality, both specific and relative humidity errors are coupled to other variables in a variety of complicated, flow-dependent ways, and neither is well described by Gaussian distributions (Gaussianity is necessary in the usual implementation of VAR; see Part I). The use of relative humidity as a moisture parameter induces δT–δq covariances, which are inappropriate away from saturation (see Section 3.2). In models with a cold bias, this problem can lead to an accumulation of moisture in the stratosphere (Dee and da Silva, 2003). Because the stratosphere is lacking observations of moisture, relative humidity will be conserved in the assimilation. Temperature observations will try to correct for the cold bias by heating, and so specific humidity will increase to conserve relative humidity; see (23). The use of specific humidity as a moisture parameter, however, can lead to damaging extrapolation errors (Dee and da Silva, 2003). These arise because of the large vertical gradients of specific humidity, allowing increments in regions of high specific humidity to spread and swamp the analysis in regions of low specific

DOI: 10.1002/qj

humidity via the effect of background error correlation (Part I).

Alternative moisture control parameters are under investigation. The Met Office is looking at a new moisture parameter (e.g. Lahoz and Geer, 2003; Lahoz *et al.*, 2006) that has the following properties: (1) it is coupled to other variables in a realistic fashion; (2) it has the properties of relative humidity under conditions close to saturation only; (3) it does not have large vertical gradients; (4) it obeys Gaussian statistics. Some of these points have been addressed previously; for example, Dee and da Silva (2003) have looked at using pseudo-relative humidity (defined as specific humidity divided by the background state's saturated specific humidity) as the moisture parameter. This can help reduce extrapolation errors as it is more homogeneous than specific humidity, and does not suffer the anomalous δT–δq correlations associated relative humidity. Centres that use pseudo-relative humidity are shown in Table I. Work by Hólm *et al.* (2002) shows that a moisture PDF can be constructed that is more Gaussian-like by a judicious choice of control parameter.

The coupling of moisture to other variables can be modelled in the same way as for pressure by introducing an 'unbalanced humidity' control parameter, $\tilde{\chi}_4 = \delta\tilde{q}_\mathrm{u}$ (i.e. the component of humidity that is not coupled to other control parameters; Lorenc *et al.*, 2003). The specific humidity increment can be written with contributions from temperature and pressure (found from the third and fourth lines of (21), but only when close to saturation), vertical motion (found from the linearized omega equation); and the new control parameter.

$$\delta q = \left\{ \begin{array}{cc} \frac{\partial q_{0\mathrm{sat}}}{\partial p}\delta p + \frac{\partial q_{0\mathrm{sat}}}{\partial T}\delta T & \text{(close to saturation)} \\ 0 & \text{(otherwise)} \end{array} \right\} + \frac{\partial q_0}{\partial z}\Delta t \delta w + \delta\tilde{q}_\mathrm{u}, \quad (40)$$

(cf. last two terms of (23) with (18)). The vertical derivative term is due to vertical advection of moisture by the vertical wind increment, δw, with an advective time-scale of Δt (Lorenc *et al.*, 2003). See Table III, row 4.

7.2. Developments to the spatial transforms

7.2.1. Modelling flow dependency, inhomogeneity and anisotropy with distorted grids in VAR

In Section 4, we demonstrated how CVTs can be used to model static, homogeneous and isotropic correlation functions. Here, we look at developments of the spatial transform via the use of distorted grids, which relax these approximations.

Riishøjgaard (1998) proposed a form of the univariate spatial background error covariance matrix whose structure functions resemble the background state itself. Elements of Riishøjgaard's **B**-matrix are

$$B_{ij} = \sigma_i \rho(r_{ij}) \gamma(|x_i^\mathrm{b} - x_j^\mathrm{b}|)\sigma_j, \quad (41)$$

where i and j label grid points, σ_i is the standard deviation of element x_i, ρ is an isotropic horizontal correlation function that decays with increasing distance, r_{ij}, between these points, and γ is a function that decays with increasing difference in magnitude, $|x_i^\mathrm{b} - x_j^\mathrm{b}|$, between the background values at these points. The basis of this correlation model is the belief that error correlations are high between nearby points that have similar background values. Riishøjgaard's method is easy to use in a data assimilation system that uses the **B**-matrix explicitly, such as optimal interpolation, but it is more difficult to apply in VAR, which would require the square root of the matrix comprising elements (41), and its inverse.

Another approach, which builds on this idea, is to still model homogeneous and isotropic error correlations, but on a grid that is distorted. This idea is used by Stajner, Riishøjgaard and Rood (2001) and more recently by Segers *et al.* (2005). The latter work introduces matrix elements of the form

$$B_{ij} = \sigma_i \gamma(|\tau(r_i) - \tau(r_j)|)\sigma_j, \quad (42)$$

(cf. (41)). Here, r_i is the position vector of grid point i in the real grid, which is displaced to a new position $\tau(r_i)$. The correlation function γ is homogeneous and isotropic in this transformed space. The properties of inhomogeneity and anisotropy are then transferred to the problem of determining the vector transformation, τ, which can, in principle, be position- and flow-dependent.

Segers *et al.* (2005) used this approach to produce an explicit forecast error covariance matrix for ozone data assimilation in a Kalman filter-like system, but it can be used also for general meteorological variables in a CVT context in VAR. This can be done by modifying the CVT to perform an additional grid transformation. The spatial transform as it appears in (13) and (14) becomes

$$\mathbf{B}_\mathrm{s}^{1/2} = \mathbf{L}_\tau \hat{\mathbf{B}}_\mathrm{s}^{1/2}, \quad (43)$$

$$\mathbf{B}_\mathrm{s}^{-1/2} = \hat{\mathbf{B}}_\mathrm{s}^{-1/2} \mathbf{L}_\tau^{-1}, \quad (44)$$

and the adjoint of (43) is also needed to work out the gradient of the cost function in (5). Operator $\mathbf{L}_\tau^{-1}$ is a matrix representation of the transform τ, which distorts positions from the real grid to the transformed grid, $\mathbf{L}_\tau$ recovers the real grid positions, and $\hat{\mathbf{B}}_\mathrm{s}^{1/2}$ and $\hat{\mathbf{B}}_\mathrm{s}^{-1/2}$ are the simple spatial transforms already introduced, but operating on the distorted grid. This formulation would work directly in physical space when used in a Met Office-like system, but would need to be expressed in spectral space when used in a ECMWF-like system, because in the latter case the spatial transform operates in a spectral representation. The implied covariance matrix for a specific parameter is from (9) and (43)

$$\mathbf{B}_\mathrm{s}^\mathrm{ic} = \mathbf{B}_\mathrm{s}^{1/2} \mathbf{B}_\mathrm{s}^{\mathrm{T}/2} = \mathbf{L}_\tau \hat{\mathbf{B}}_\mathrm{s}^\mathrm{ic} \mathbf{L}_\tau^\mathrm{T}, \quad (45)$$

DOI: 10.1002/qj

where $\hat{\mathbf{B}}_s^{ic}$ is homogeneous and isotropic (see (28) or (33)), but $\mathbf{B}_s^{ic}$ is not.

Segers *et al.* (2005) determined their transformation such that the correlations on the real grid matched closely the correlations found by analysing forecast differences. A special distorted grid has been used by Desroziers and Lafore (1993) and Desroziers (1997) who used the geostrophic coordinate transform (GCT) of semigeostrophic theory to distort the grid. A field of perturbations (specifically in VAR one of the control parameters, $\tilde{\chi}_i$), which is expressed as a function of the Cartesian coordinates x_R, y_R and z_R, can be distorted by the GCT to so-called geostrophic coordinates, x_G, y_G and z_G:

$$x_G = x_R + f^{-1}v_g, \quad y_G = y_R - f^{-1}u_g, \quad z_G = z_R. \tag{46}$$

Here, u_g and v_g are the x and y components of the geostrophic wind, respectively, which in practice can be derived from the background field. The GCT going from real (x_R, y_R, z_R) to geostrophic (x_G, y_G, z_G) coordinates is the application of $\mathbf{L}_\tau^{-1}$. The displacement in each horizontal direction is the distance moved by a parcel in one inertial period by the geostrophic wind in the orthogonal direction. The gridded values of $\tilde{\chi}_i$ are each preserved from their positions on the real grid to their transformed positions on the geostrophic grid. However, as the real and geostrophic grids are both regular grids, some interpolation is required during the transformation.

In semigeostrophic theory (Hoskins and Bretherton, 1972; Hoskins, 1975), the semigeostrophic equations of motion simplify after they have been transformed by the GCT, where the Lagrangian time derivatives of positions x_G and y_G are just u_g and v_g, respectively. On the geostrophic grid, many features such as fronts disappear. This is the basis of the GCT, where it is reasonably assumed that forecast errors on the geostrophic grid are more homogeneous and isotropic than those on the real grid.

The GCT technique has been described in the context of the inverse transform (44), but in the assimilation it is used as (43), that is, transforming from the geostrophic grid, where correlations are taken to be homogeneous and isotropic, to the real grid (see Table III, row 5).

7.2.2. *Wavelet formulation of forecast errors*

The possible ways that correlations can be localized by a model of the **B**-matrix has been explored in recent years (e.g. Buehner and Charron, 2007). Consider a model of the **B**-matrix that allows vertical covariances to vary with horizontal position, but embodies no covariances between horizontal points. This is a scheme that uses spatial localization in the horizontal. However, consider a different model of the **B**-matrix that allows vertical covariances to vary with horizontal wave number, but has no covariances between different wave numbers. This is a spectral localization. It will capture the way that vertical covariances vary with horizontal scale, but will not capture the position dependence, or inhomogeneities. Real forecast errors often have strong position and scale dependences of vertical covariances (see Figure 10 of Part I). Wavelet-based methods of modelling the **B**-matrix are intermediate to these extremes (e.g. Deckmyn and Berre, 2005) and so are useful in modelling more general aspects of forecast error covariances.

A wavelet may be regarded as a wave-like structure that can have (simultaneously) a characteristic scale and position identified with it. A wavelet basis comprises members that can each be characterized in this way. Fisher (2003) shows a set of simple wavelet functions, $\psi_j(r)$, and their spectral transforms, $\hat{\psi}_j(n)$, which demonstrates this property (n is the total horizontal wave number, r is the horizontal distance and j is the wavelet index). This simultaneous scale and position property of wavelets makes them attractive for modelling scale- and position-dependent aspects of covariances. A wavelet $\psi_j(r)$ acts as a bandpass filter when convolved with another function. The range of wave numbers (the 'band') that $\psi_j(r)$ allows through depends upon j, from the largest scales (e.g. $j = 1$) to the smallest (e.g. $j = K$), where K is the number of wavelets considered. These functions can be built into a CVT in the following way (Fisher, 2003, 2004)

$$\mathbf{B}_s^{1/2}\chi = \mathbf{S}^{-\mathrm{T}}\mathbf{V}^{1/2}\sum_{j=1}^{K}\psi_j(r)\otimes\mathbf{C}_j^{1/2}\chi_j. \tag{47}$$

This replaces (26) for the ECMWF and is described below (we have dropped the parameter index from all symbols). Similar transforms have been considered for other systems, for example, in the Météo-France Aladin system by Deckmyn and Berre (2005) and the Met Office system by Andrews and Lorenc (personal communication) and Bannister (2007). Equation (47) is more complicated than the standard ECMWF spatial transform. Instead of one control vector per parameter, there are now K subvectors, $\chi = (\chi_1, \chi_2, \ldots, \chi_K)^{\mathrm{T}}$, where χ_j is associated with wavelet j (there is one such set for each parameter). In (26), χ represents a parameter's field as a function of wave number and vertical mode, but in (47), each χ_j is a function of longitude, latitude and height (the scheme can be modified to use vertical mode instead of height). Each subvector has its own vertical transform, $\mathbf{C}_j^{1/2} = \mathbf{E}_j\mathbf{D}_j^{1/2}\mathbf{E}_j^{\mathrm{T}}$, which is the symmetric square root of the vertical covariance matrix at each position for wavelet j. This square root is of the general form of (11): $\mathbf{E}_j^{\mathrm{T}}$ projects each vertical column in χ_j onto the vertical eigenvectors local to each position, the diagonal $\mathbf{D}_j^{1/2}$ multiplies the result by the square root of the eigenvalues, and $\mathbf{E}_j$ projects back to a height representation. In (26), there is a different vertical transform for each wave number, but in (47) there is a different vertical transform for each horizontal position. If it were not for the presence of the wavelets, this formulation of the vertical transform would prohibit the prescription of scale dependences of the vertical covariances, but in (47) these appear through the j dependence of $\mathbf{C}_j^{1/2}$. The field

DOI: 10.1002/qj

resulting from action with each $\mathbf{C}_j^{1/2}$ is convolved with ψ_j (thus performing the bandpass filtering). The diagonal matrix $\mathbf{V}^{1/2}$ contains the standard deviations of the grid-point parameters and is the same matrix used in (26). $\mathbf{S}^{-\mathrm{T}}$ transforms the resulting field to spectral space, as required in the ECMWF system.

The implied point-to-point correlations, $C^{\mathrm{ic}}(r', z'; r, z)$, can be found by application of (9) to the correlation-only part of (47) (i.e. $\sum_j \psi_j \otimes \mathbf{C}_j^{1/2}$). Unlike the result for the basic ECMWF scheme (29), horizontal correlations here allow inhomogeneity. The implied correlation between horizontal position r and height z and another point at r', z' is

$$C^{\mathrm{ic}}(r', z'; r, z) = \sum_{j=1}^{K} \int \mathrm{d}r'' \psi_j(r' - r'') \times \psi_j(r'' - r) C_j(z', z; r''), \quad (48)$$

which has the general property that correlations are no longer a function of the separation between a given two points, $r' - r$. The method can also be used to model anisotropy (Deckmyn and Berre, 2005). Pannekoucke, Berre and Desroziers (2007) show additionally that background error correlations modelled with such a 'wavelet diagonal' approach can filter sampling noise introduced when representing correlations with a small number of ensemble members (filtering has been dealt with before using the Schur product; Houtekamer and Mitchell, 2001). Transform (47) is now operational in the ECMWF system.

The number of bands, K, needs to be chosen. Too few bands will result in inadequate spectral resolution, but too many bands may result in the loss of spatial localization. There is thus an interplay between the effective spectral and spatial resolutions (Bannister, 2007). See Table III, row 6.

7.2.3. *Recursive filters*

Recursive filters have been considered for modelling the spatial parts of the **B**-matrix in the horizontal (Lorenc, 1997), and can be posed in the form of a CVT. A simple formulation of recursive filter to model homogeneous covariances is discussed below, which can be applied as an alternative to the spectral approach (i.e. by assuming that covariances between different wave numbers are zero) as used e.g. at the ECMWF; (see Table I). The potential of the method shown here is that it can be adapted efficiently to model inhomogeneities and anisotropies. Recursive filters are used currently in the NCAR mesoscale 3d-VAR system (Barker *et al.*, 2004) and the NCEP real-time mesoscale analysis system (de Pondeca *et al.*, 2007).

The basic principle can be understood by considering the convolution (in grid space) of a field with a Gaussian-shaped kernel (Purser *et al.*, 2003a). Such a convolution is equivalent to acting on a state with a covariance matrix with homogeneous and Gaussian-shaped structure functions. Let $\mathbf{B}_1$ be such a covariance matrix, defined in one dimension for simplicity. A recursive filter usually works directly in grid space but our explanation starts with its representation in spectral space, $\hat{\mathbf{B}}_1$. Let $\delta\mathbf{f}$ and $\delta\mathbf{g}$ be the vector representations of functions $\delta f(r)$ and $\delta g(r)$, respectively, in grid space and let $\delta\hat{\mathbf{f}}$ and $\delta\hat{\mathbf{g}}$ be their vector representations of their Fourier transforms, $\delta\hat{f}(k)$ and $\delta\hat{g}(k)$ (where r is position and k is wave number). By the convolution theorem of Fourier transforms, $\hat{\mathbf{B}}_1$ is the diagonal matrix whose diagonal elements comprise the Fourier transform of the Gaussian-shaped structure function. The matrix expression $\delta\hat{\mathbf{f}} = \hat{\mathbf{B}}_1 \delta\hat{\mathbf{g}}$ and its inverse, $\delta\hat{\mathbf{g}} = \hat{\mathbf{B}}^{-1}\delta\hat{\mathbf{f}}$ are equivalent to

$$\delta\hat{f}(k) = b_0 \exp(-a^2k^2/2)\delta\hat{g}(k), \quad (49a)$$

$$\delta\hat{g}(k) = b_0^{-1} \exp(a^2k^2/2)\delta\hat{f}(k), \quad (49b)$$

where $b_0 \exp(-a^2k^2/2)$ is diagonal element k of $\hat{\mathbf{B}}_1$, a is the characteristic length-scale of the Gaussian and b_0 is a constant. To transform to grid space, first expand the exponential as a Taylor series to order n in (49b)

$$\delta\hat{g}(k) \approx b_0^{-1} \sum_{i=1}^{n} c_i \left(\frac{a^2k^2}{2}\right)^i \delta\hat{f}(k), \quad (50)$$

where $c_i = b_0/i!$. In grid space, (50) is

$$\delta g(r) \approx b_0^{-1} \sum_{i=1}^{n} c_i (-1)^i \left(\frac{a^2}{2}\right)^i \frac{\mathrm{d}^{(2i)}}{\mathrm{d}r^{(2i)}} \delta f(r) \quad (51)$$

(it is more straightforward to prove (50) from (51)). For use as a CVT in VAR, $\mathbf{B}_1^{1/2}$ and its adjoint are needed. The operator acting upon $\delta f(r)$ in (51) is an approximation of $\mathbf{B}_1^{-1}$ and can be written as a Cholesky decomposition (Purser *et al.*, 2003a), which comprises a lower triangular matrix, **L**, and an upper triangular matrix, **U** where $\mathbf{U} = \mathbf{L}^{\mathrm{T}}$, giving $\mathbf{B}_1^{-1} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$. Taking the inverse of this gives $\mathbf{B}_1 = \mathbf{U}^{-1}\mathbf{U}^{-\mathrm{T}}$, or in two steps,

$$\delta\mathbf{q} = \mathbf{U}^{-\mathrm{T}}\delta\mathbf{g} \quad \text{and} \quad \delta\mathbf{f} = \mathbf{U}^{-1}\delta\mathbf{q}, \quad (52)$$

where $\delta\mathbf{q}$ is an intermediate state. Because of the triangular structure of **U**, each matrix is easy to invert. On an infinite domain

$$\delta q(r_i) = \beta\delta g(r_i) + \sum_{j=1}^{n} \alpha_j \delta q(r_i - r_j), \ i \text{ increasing}, \quad (53a)$$

$$\delta f(r_i) = \beta\delta q(r_i) + \sum_{j=1}^{n} \alpha_j \delta f(r_i + r_j), \ i \text{ decreasing}, \quad (53b)$$

where $\beta = 1/U_{ii}$ and $\alpha_j = -U_{i,i+j}/U_{ii}$. In the first step i is increasing and in the second step it is decreasing. The appearance of the solution vector on both sides of (53a) and (53b) gives these equations their recursive nature. These equations have been written with a fixed upper index, n, on the summations to reflect the order of the

DOI: 10.1002/qj

filter, where larger values of n give better approximations of the Gaussian filter. Lorenc (1992) discusses a first-order recursive filter of this type and states that a large number of passes with a first-order filter equivalently gives a Gaussian filter. Looking at the structure $\mathbf{B}_1 = \mathbf{U}^{-1}\mathbf{U}^{-\mathrm{T}}$ reveals that $\mathbf{U}^{-1}$ (53b) may be treated as the CVT for approximate Gaussian-shaped horizontal error covariances and $\mathbf{U}^{-\mathrm{T}}$ (53a) is its adjoint.

This problem may be generalized to work in two (Purser *et al.*, 2003a) and three (Wu *et al.*, 2002) dimensions, and to include the effects of inhomogeneity and anisotropy (Purser *et al.*, 2003b). A more complete analysis of the problem than that presented here considers properly the discrete nature of the underlying grid (Purser *et al.*, 2003a). The method may be extended to the sphere, but difficulties arise especially at the poles because of the use of a non-uniform grid (Wu *et al.*, 2002). A more complete review of recursive filters is provided by Raymond and Garder (1991). See Table III, row 7.

7.2.4. *Diffusion operators*

The action of a covariance matrix on a state can be shown to be equivalent to the integration of a carefully constructed diffusion equation over a finite time period using the same state as the initial conditions (Weaver and Courtier, 2001). The correlation length-scales emerge as $\sqrt{2\kappa T}$ where κ is the diffusion coefficient and T is the integration period. By allowing κ to vary in space, it is possible to model inhomogeneity. It is also possible to prescribe the shape of the correlation functions and to build anisotropy into the scheme. This equivalence between a covariance and the solution of a diffusion equation has been exploited in the ocean variational data assimilation schemes of Derber and Rosati (1989) and Weaver and Courtier (2001). The method is suited to oceanic data assimilation as it can account naturally for the coastal boundary conditions found in these problems. Like the recursive filter, however, it requires factorization into a square-root form before it can be used as a CVT (see Table III, row 8).

7.3. Other developments to the control variable transforms to allow flow dependence

The issue of flow dependence in the forecast error covariance matrix has focused recently on the ensemble Kalman filter (EnKF; e.g. Evensen, 2003). The EnKF is considered to be the main competitor to the variational method. Despite its nice theoretical properties, and its ease of implementation, there remain many hurdles to overcome before the EnKF can be used reliably (e.g. Ehrendorfer, 2007) and so VAR remains the workhorse of operational data assimilation. There is therefore much interest in developing VAR to deal with flow dependence. The developments already mentioned in Sections 7.1 and 7.2 are concerned with the parameter and spatial parts of the CVT, but further developments can be made to introduce flow dependence in $\mathbf{B}$ by making changes to the structure of the CVT. Two important examples are described below: the reduced rank Kalman filter and the 'errors of the day' scheme. Each method gives special treatment to the background error covariances of a dynamically active subspace of the system.

7.3.1. *The reduced rank Kalman filter*

The reduced rank Kalman filter (RRKF) (also known as the simplified Kalman filter) is an attempt to treat a dynamically active part of the background error in an explicit way (akin to the way that the Kalman filter treats forecast error covariances), while remaining in a VAR framework. The scheme is outlined briefly in Rabier *et al.* (1997), and more fully by Fisher (1998) and Beck and Ehrendorfer (2005).

The RRKF allows an explicit propagation of part of the analysis error covariance from the previous cycle (at time $t = -T$) to the time of the background state in the current cycle (at time $t = 0$). The part of the analysis error covariance chosen is that associated with the leading K Hessian singular vectors, $\delta\mathbf{x}_k(-T)$, $1 \leq k \leq K$. These singular vectors are those perturbations at $t = -T$ that have maximum growth over a specified time period as measured by the inverse Hessian (Barkmeijer, van Gijzen and Bouttier, 1998; Barkmeijer, Buizza and Palmer, 1999). Their propagation to $t = 0$ (i.e. $\mathbf{s}_k = \mathbf{M}_{0\leftarrow -T}\delta\mathbf{x}_k(-T)$) defines the subspace for which background errors are treated explicitly. An increment at $t = 0$, $\delta\mathbf{x}$, then has a contribution that is within this subspace, $\delta\mathbf{x}_s$, and a part that is orthogonal to it, $\delta\bar{\mathbf{x}}_s$:

$$\delta\mathbf{x} = \delta\mathbf{x}_s + \delta\bar{\mathbf{x}}_s. \tag{54}$$

Making this partition in the background part of the cost function (1) gives three terms

$$\begin{aligned} J_\mathrm{b}(\delta\mathbf{x}, \mathbf{x}^\mathrm{g}) = {} & \frac{1}{2}(\delta\mathbf{x}_s - \delta\mathbf{x}_s^\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\delta\mathbf{x}_s - \delta\mathbf{x}_s^\mathrm{b}) \\ & + (\delta\bar{\mathbf{x}}_s - \delta\bar{\mathbf{x}}_s^\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\delta\mathbf{x}_s - \delta\mathbf{x}_s^\mathrm{b}) \\ & + \frac{1}{2}(\delta\bar{\mathbf{x}}_s - \delta\bar{\mathbf{x}}_s^\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\delta\bar{\mathbf{x}}_s - \delta\bar{\mathbf{x}}_s^\mathrm{b}), \end{aligned} \tag{55}$$

where increments are with respect to a guess state, $\mathbf{x}^\mathrm{g}$, and superscript 'b' refers to the background. The first term of (55) involves only the special subspace, the third term involves only the remaining part of the state space and the second term involves a mixture. The RRKF is constructed by imposing a flow-dependent error covariance matrix for the first two terms ($\mathbf{B} \rightarrow \mathbf{P}^\mathrm{f}$), while keeping the usual static matrix for the third term as follows:

$$\begin{aligned} J_\mathrm{b}(\delta\mathbf{x}, \mathbf{x}^\mathrm{g}) \rightarrow {} & \frac{1}{2}(\delta\mathbf{x}_s - \delta\mathbf{x}_s^\mathrm{b})^\mathrm{T}\mathbf{P}^{\mathrm{f}\,-1}(\delta\mathbf{x}_s - \delta\mathbf{x}_s^\mathrm{b}) \\ & + \alpha(\delta\bar{\mathbf{x}}_s - \delta\bar{\mathbf{x}}_s^\mathrm{b})^\mathrm{T}\mathbf{P}^{\mathrm{f}\,-1}(\delta\mathbf{x}_s - \delta\mathbf{x}_s^\mathrm{b}) \\ & + \frac{1}{2}(\delta\bar{\mathbf{x}}_s - \delta\bar{\mathbf{x}}_s^\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\delta\bar{\mathbf{x}}_s - \delta\bar{\mathbf{x}}_s^\mathrm{b}). \end{aligned} \tag{56}$$

The factor α in the second term was added by Fisher (1998) to help keep the cost function convex. The hypothesis of the RRKF is that the small subspace $\delta\mathbf{x}_s$ will

DOI: 10.1002/qj

contribute significantly to the background error covariance at $t = 0$ and so concentrating effort on representing $\mathbf{P}^{\mathrm{f}}$ realistically will improve the assimilation. The remaining problems are as follows: (1) how to modify the CVT to allow the subspace to be treated separately and (2) how to determine $\mathbf{P}^{\mathrm{f}}$ from available information.

To deal with (1), VAR's standard CVT, $\mathbf{B}_0^{1/2}$ in (8), can be modified with an extra orthogonal transformation,

$$\delta\mathbf{x} = \mathbf{K}_{\mathrm{p}}\left[\beta_0\begin{pmatrix}\mathbf{B}_{\mathrm{s},\tilde{\chi}_1}^{1/2} & 0 & 0\\ 0 & \mathbf{B}_{\mathrm{s},\tilde{\chi}_2}^{1/2} & 0\\ 0 & 0 & \mathbf{B}_{\mathrm{s},\tilde{\chi}_3}^{1/2}\end{pmatrix}\begin{pmatrix}\chi_1\\ \chi_2\\ \chi_3\end{pmatrix} + \beta_1\sum_{k=1}^{p}\begin{pmatrix}\mathbf{V}_1^k(\mathbf{B}_{\mathrm{s},1}^k)^{1/2} & 0 & 0\\ 0 & \mathbf{V}_2^k(\mathbf{B}_{\mathrm{s},2}^k)^{1/2} & 0\\ 0 & 0 & \mathbf{V}_3^k(\mathbf{B}_{\mathrm{s},3}^k)^{1/2}\end{pmatrix}\begin{pmatrix}\mathrm{X}_1^k\\ \mathrm{X}_2^k\\ \mathrm{X}_3^k\end{pmatrix}\right]. \quad (61)$$

$\mathbf{X}$, as follows

$$\delta\mathbf{x} = \mathbf{B}_0^{1/2}\mathbf{X}\chi, \quad (57)$$

where the control variable, χ, is the same as in the standard VAR (because of the added flow dependency in (56), however, the implied covariance matrix is no longer equal to $\mathbf{B}_0^{1/2}\mathbf{B}_0^{\mathrm{T}/2}$). In the RRKF, $\mathbf{X}$ is designed to have the following properties:

- $\mathbf{X}$ is orthogonal, $\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{I}$;
- $\mathbf{X}^{\mathrm{T}}$ acting on a vector in the subspace $\mathbf{B}_0^{-1/2}\delta\mathbf{x}_s$ gives a vector that can be non-zero only in the first K elements

$$\mathbf{X}^{\mathrm{T}}\mathbf{B}_0^{-1/2}\delta\mathbf{x}_s = (a_1, \ldots, a_K, 0, \ldots 0)^{\mathrm{T}}, \quad (58)$$

- $\mathbf{X}^{\mathrm{T}}$ acting on a vector in the remaining subspace $\mathbf{B}_0^{-1/2}\delta\bar{\mathbf{x}}_s$ gives a vector that can be non-zero only in the remaining elements.

Property (58) ensures that the part of the state space associated with the flow-dependent error covariances is represented entirely by the first K elements of the control vector and the remaining part with static error covariances is represented by the rest. These properties are possible by constructing $\mathbf{X}^{\mathrm{T}}$ (and consequently $\mathbf{X}$) as a sequence of Householder transformations (Fisher, 1998; also see Bannister, 2006, for details). Substituting (57) into (56) gives

$$\begin{aligned}J_{\mathrm{b}}(\delta\mathbf{x}, \mathbf{x}^{\mathrm{g}}) =& \frac{1}{2}(\chi_s - \chi_s^{\mathrm{b}})^{\mathrm{T}}\mathbf{P}_{\chi}^{\mathrm{f}\,-1}(\chi_s - \chi_s^{\mathrm{b}})\\ &+ \alpha(\bar{\chi}_s - \bar{\chi}_s^{\mathrm{b}})^{\mathrm{T}}\mathbf{P}_{\chi}^{\mathrm{f}\,-1}(\chi_s - \chi_s^{\mathrm{b}})\\ &+ \frac{1}{2}(\bar{\chi}_s - \bar{\chi}_s^{\mathrm{b}})^{\mathrm{T}}(\bar{\chi}_s - \bar{\chi}_s^{\mathrm{b}}), \end{aligned} \quad (59)$$

where, as in standard VAR, $\mathbf{B}_0^{1/2}$ is assumed to be an accurate square root of the static $\mathbf{B}$-matrix giving $\mathbf{X}^{\mathrm{T}}\mathbf{B}_0^{\mathrm{T}/2}\mathbf{B}^{-1}\mathbf{B}_0^{1/2}\mathbf{X} \approx \mathbf{I}$. For (59), the following has been defined:

$$\mathbf{P}_{\chi}^{\mathrm{f}\,-1} = \mathbf{X}^{\mathrm{T}}\mathbf{B}_0^{\mathrm{T}/2}\mathbf{P}^{\mathrm{f}-1}\mathbf{B}_0^{1/2}\mathbf{X}. \quad (60)$$

Control vector contributions χ_s and χ_s^{b} comprise K components and $\bar{\chi}_s$ and $\bar{\chi}_s^{\mathrm{b}}$ comprise $n - K$ components (for the state vector having n-dimensions). Problem (2) is dealt with by first noting that because $\mathbf{P}_{\chi}^{\mathrm{f}\,-1}$ acts only on vectors of K components, then only the first K columns of $\mathbf{P}_{\chi}^{\mathrm{f}\,-1}$ need be known. In trials of the RRKF, K is relatively small (10–25; Fisher, 1998; Fisher and Andersson, 2001) and so $\mathbf{P}_{\chi}^{\mathrm{f}\,-1}$ may be stored comfortably as an explicit matrix. Fisher (1998) shows how these K columns can be calculated from the singular vector calculation. No account is taken of model error in the propagation of covariance information from the previous analysis to the current background, although Beck and Ehrendorfer (2005) show how model error can be added.

There has been mixed success with the RRKF. Hello and Bouttier (2001) experimented with a similar method with $K = 1$ in the Météo-France Arpège 3d-VAR system to see if it could be a cheaper alternative to 4d-VAR. They report that it performs better than the bare 3d-VAR in most cases. A number of studies have been made at the ECMWF with 4d-VAR modified by the RRKF, using 10–25 singular vectors to define the subspace, although Ehrendorfer and Bouttier (1998) suggest that more (e.g. 100) vectors may be needed. Fisher (1998) reports a small positive impact of the RRKF verses bare 4d-VAR in 6–10 day forecasts, but a later study by Fisher and Andersson (2001) reports that the RRKF has only a neutral impact over a number of cases (see Table III, row 9).

7.3.2. *Errors of the day*

Another approach that incorporates flow-dependent information into $\mathbf{B}$ introduces contributions to the structure functions that are explicitly governed by the dynamically active modes of the system. So-called 'bred modes' found by the nonlinear error breeding technique can be used to define the dynamically active modes for this purpose (Kalnay *et al.*, 1997).

Error breeding is a method used to generate rapidly growing modes that evolve under the governing nonlinear forecast model (Toth and Kalnay, 1993). Because of the way that they are constructed, these modes are called 'bred modes', $\mathbf{v}^k$ (where k is a mode index), and they will change from day to day. $\mathbf{v}^k$ may be interpreted as the leading error patterns, characteristic of the structure of forecast errors (e.g. Corazza *et al.*, 2003). In this context they are also called the forecast 'errors of the day'. Toth and Kalnay (1997) argue that growing modes are the most important component of forecast error. The error breeding technique is straightforward to implement in principle,

DOI: 10.1002/qj

and involves repeated application of the forecast model to small perturbations followed by normalization (Toth and Kalnay, 1993). Because the generation of bred modes is inexpensive, it is anticipated that building bred modes into the **B**-matrix may represent an effective and cheap alternative to data assimilation methods that propagate forecast errors explicitly.

Bred modes were proposed in data assimilation studies before the use of VAR, but not in a way that directly involves the **B**-matrix. Kalnay and Toth (1994) used bred modes to perform a pre-analysis fit to observations. This procedure allows the background state to be adjusted in the subspace spanned by the bred modes, which adapts with the synoptic situation. The static **B**-matrix used in the main assimilation may then be more appropriate to the adjusted background than the unadjusted one, as the state-dependent part of the errors will be reduced by the pre-analysis. A different strategy was adopted by Pu *et al.* (1997). They used bred modes to make a judicious modification of observation errors, which were lowered in regions of large bred mode amplitude. In these regions, this procedure increases the importance of observations relative to the background. Both studies report positive impacts.

Barker and Lorenc (2005) outline a scheme that modifies the implied **B**-matrix in VAR by extending the standard form of the Met Office CVT with extra contributions from the bred vectors. In their scheme, which uses p bred vectors per assimilation cycle, the standard control vector is augmented with p new subvectors. Their CVT has the form of (61) (this example uses three control parameters).

The original CVT is the first term, as (15), and the new (remaining) part of the transform is described below in parts. Each original subvector, χ_i, is supplemented with new subvectors, denoted by X_i^k, associated with the bred modes ($1 \leq k \leq p$). There is one new subvector per parameter and per bred mode, which is why X_i^k has two indices, k and i. Components of the new control vector

$$\chi = (\chi_1, \chi_2, \chi_3, \mathrm{X}_1^1, \mathrm{X}_2^1, \mathrm{X}_3^1, \mathrm{X}_1^2, \mathrm{X}_2^2, \mathrm{X}_3^2, \ldots), \qquad (62)$$

are taken to be mutually uncorrelated and with unit variance. X_i^k may exist in the same space as χ_i (e.g. a function of horizontal wave number and vertical mode), but for efficiency their resolution may be lower (e.g. initial trials at the Met Office were performed with X_i^k of only T10 resolution, with only one vertical mode, and one bred mode, $p = 1$). The spatial operators $(\mathbf{B}_{\mathrm{s},i}^k)^{1/2}$ transform the extra control vectors to grid space, and play a similar role to $\mathbf{B}_{\mathrm{s},\tilde{\chi}_i}^{1/2}$ for the standard control vectors (Section 4). The diagonal elements of the diagonal matrices $\mathbf{V}_{\chi_i}^k$ comprise the normalized bred vectors $\mathbf{v}^k$. The product $\mathbf{V}_i^k(\mathbf{B}_{\mathrm{s},i}^k)^{1/2}\mathrm{X}_i^k$ is essentially an element-by-element ('Schur' or 'Hadamard') product (e.g. Gaspari and Cohn, 1999; Lorenc, 2003b) between $\mathbf{v}^k$ and $(\mathbf{B}_{\mathrm{s},i}^k)^{1/2}\mathrm{X}_i^k$, which gives this transform its flow dependence. The original and new terms of the CVT are weighted by coefficients β_0 and β_1, respectively, which can be tuned, but must satisfy $\beta_0^2 + \beta_1^2 = 1$.

Minimization of the cost function can be performed with respect to all components of (62). By defining χ as in (62), (61) may be written as a composite transform, $\delta\mathbf{x} = \mathbf{B}_0^{1/2}\chi$ as in (8), but with a CVT that is more complicated than before. Even though this CVT is not square, it remains a valid square root. The CVT allows an analytical form of the implied covariances to be derived using (9). The point-to-point covariances of the single parameter $\tilde{\chi}_i$ (see Section 2.3) has the form

$$\mathbf{B}_{\chi_i}^{\mathrm{ic}} = \beta_0^2\mathbf{B}_{\mathrm{s},\chi_i} + \beta_1^2\sum_{k=1}^{p}\mathbf{V}_i^k\mathbf{B}_{\mathrm{s,i}}^k\mathbf{V}_i^k. \qquad (63)$$

The standard spatial covariances $\mathbf{B}_{s,\chi_i}$ and the bred mode spatial covariances $\mathbf{B}_{s,i}^k$ are not flow-dependent, but flow dependency is brought in through the bred modes in (63). The right-hand term has some similarity to the Riishøjgaard flow-dependent background error covariance model (Riishøjgaard, 1998). In Riishøjgaard's model (41), the prescribed covariances are modulated by a function that is the difference between the background state at two positions, but in (63) they are modulated by the bred vector values at these positions. Semple (2002) has trialled (61) with an integrated breeding scheme at the Met Office, but early tests have found the impact to be neutral (see Table III, row 10).

8. Concluding remarks

This article is an introduction to and a review of the CVT method of modelling the **B**-matrix in variational assimilation. The **B**-matrix affects the performance of the assimilation (reviewed in Part I) and so it is important to use a **B**-matrix that is a realistic representation of the actual forecast error covariances. Because of its prohibitive size, it is not possible to use the explicit form of the **B**-matrix, but we demonstrate how an approximation can be constructed by expressing the cost function in terms of new control variables whose background errors are assumed to be mutually uncorrelated. A basic form of the CVT is a sequence of spatial and parameter transforms, which recover the model variables from the control variables, and in the process give rise to implied correlations between the model variables. There is a degree of flexibility in the design of CVTs and consequently a wide range of variations of CVTs have been used by different centres over recent years. This has led to an extensive literature on the subject. This article gives special reference to the ECMWF and Met Office formulations as illustrative examples.

By assuming that background errors are static, are in hydrostatic and near geostrophic balance, that structures of control parameter errors are homogeneous in the horizontal, it is possible to model the **B**-matrix with a small fraction of the information needed to define explicitly

DOI: 10.1002/qj

the full matrix (Table II). The missing information is filled in by the above assumptions automatically by the CVT. It is the feasibility of the CVT transform method that has allowed VAR to be a pragmatic and successful method.

Some of the consequences that popular choices of CVT have on the implied structure functions are outlined, namely homogeneity, isotropy and near separability between the vertical and horizontal directions. These, and sometimes inappropriate treatments of dynamical balance and of moisture in the CVT, can have undesirable consequences on the assimilation. These drawbacks and the absence of adequate flow dependence in the **B**-matrix have been discussed, and a selection of contemporary developments to help solve these problems (still using the CVT methodology) have been reviewed. These include the use of state-dependent (instead of static) balance relationships (including potential vorticity and omega-equation based balance relations), geostrophic coordinate transforms, wavelets and recursive filters. It is possible also to extend the CVT method to allow VAR to make a special treatment of the most dynamically active parts of the phase space in each VAR cycle. The workings of two methods, namely the RRKF and the 'errors of the day' scheme, which were developed to bring realistic flow dependence to VAR, have been reviewed. Studies have found that these methods are not beneficial in some trials, but related methods may, however, become more important in future applications, such as in high-resolution data assimilation systems (see below). It is also possible to extend the control variable to treat model error in weak constraint 4d-VAR, but this is outside the scope of this article.

VAR is expected to remain the mainstay of data assimilation for the foreseeable future in many forecasting centres, and it will be used with increasingly higher resolution models. Météo-France, for example, have already published plans for a VAR-based high-resolution limited-area system with a model of 2.5 km grid length (Fischer *et al.*, 2005). With high-resolution models, many of the core assumptions used currently in background error covariance modelling become questionable, which raise new and important challenges in appropriate future designs of CVTs. This is particularly true with regard to the relevance of dynamical balance relationships at the 'convective' scale (otherwise known as the 'cumulus' or 'storm' scale). Convective-scale effects become visible when the model's grid size is reduced below a few km. These are expected to be described poorly by a **B**-matrix that is formulated with the set of assumptions that are often made presently (e.g. hydrostatic and geostrophic balance, isotropy, homogeneity, Gaussianity and staticness of the **B**-matrix), and may have complicated dynamics–moisture interactions. For these reasons, the **B**-matrix is expected to become highly flow-dependent. It is not yet certain if it is possible to adapt existing schemes to the convective scale in a satisfactory way, especially if small and large scales need to be analysed simultaneously (Lorenc and Payne, 2007), and so studies are required to investigate these effects and how they can be modelled effectively.

Acknowledgements

I would like to thank the many people who have influenced my understanding of data assimilation over the past few years: Phil Andrews, Sue Ballard, Paul Berrisford, Roger Brugge, Mike Cullen, Sarah Dance, Mark Dixon, Martin Ehrendorfer, John Eyre, Mike Fisher, Alan Geer, Bruce Ingleby, David Jackson, Mike Keil, William Lahoz, Amos Lawless, Andrew Lorenc, Stefano Migliorini, Alan O'Neill, Nancy Nichols, Dave Pearson, Ian Roulstone, Olaf Stiller, Hamish Struthers, Sean Swarbrick, Richard Swinbank and Matt Szyndel. Particular thanks go to Martin Ehrendorfer, Mike Fisher, Bruce Ingleby, Andrew Lorenc and Alan O'Neill for reading the draft manuscript. I would like to thank three anonymous reviewers for their valuable comments and also the authors of the work cited in this article. I gratefully acknowledge the Met Office for permission to reproduce Figure 2. The author is funded by the Natural Environment Research Council.

References

Bannister RN. 2004. On control variable transforms in the Met Office 3d and 4d Var, and a description of the proposed waveband summation transformation. DARC Internal Report, **5** (available from Data Assimilation Research Centre, Department of Meteorology, University of Reading, Reading, RG6 6BB, UK; http://darc.nerc.ac.uk/v2/internalreports.html).

Bannister RN. 2006. Collection of results on background error covariance models. DARC Internal Report, **7** (available from Data Assimilation Research Centre, Department of Meteorology, University of Reading, Reading, RG6 6BB, UK; http://darc.nerc.ac.uk/v2/internalreports.html).

Bannister RN. 2007. Can wavelets improve the representation of forecast error covariances in variational data assimilation? *Mon. Weather Rev.* **135**: 387–408.

Bannister RN. 2008. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Q. J. R. Meteorol. Soc.* **134**. DOI:10.1002/qj.339.

Bannister RN, Cullen MJP. 2006. The implementation of a PV-based leading control variable in variational data assimilation. Part 1: Transforms. DARC Internal Report, **6** (available from Data Assimilation Research Centre, Department of Meteorology, University of Reading, Reading, RG6 6BB, UK; http://darc.nerc.ac.uk/v2/internalreports.html).

Bannister RN, Katz D, Cullen MJP, Lawless, AS, Nichols NK. 2008. Modelling of forecast errors in geophysical fluid flows. *Int. J. Numer. Meth. Fluids* **56**: 1147–1153.

Barker DM, Lorenc AC. 2005. The use of synoptically-dependent background error structures in 3d Var. Var. Scientific Documentation Paper, **26** (available from Met Office, Fitzroy Road, Exeter, Devon, EX8 3PB, UK).

Barker DM, Huang W, Guo YR, Bourgeois AJ. 2003. A three-dimensional variational (3DVAR) data assimilation system for use with MM5. NCAR Technical Note **453** (available from UCAR Communications, PO Box 3000, Boulder, CO 80 307, USA).

Barker DM, Huang W, Guo YR, Bourgeois AJ, Xiao QN. 2004. A three-dimensional variational data assimilation system for MM5: Implementation and initial results. *Mon. Weather Rev.* **132**: 897–914.

Barkmeijer J, van Gijzen M, Bouttier F. 1998. Singular vectors and estimates of the analysis-error covariance metric. *Q. J. R. Meteorol. Soc.* **124**: 1695–1713.

Barkmeijer J, Buizza R, Palmer TN. 1999. 3d-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**: 2333–2351.

Bartello P, Mitchell HL. 1992. A continuous three-dimensional model of short-range forecast error covariances. *Tellus* **44A**: 217–235.

Beck A, Ehrendorfer M. 2005. Singular-vector-based covariance propagation in a quasigeostrophic assimilation system. *Mon. Weather Rev.* **133**: 1295–1310.

Berre L. 2000. Estimation of synoptic and mesoscale forecast error covariances in a limited area model. *Mon. Weather Rev.* **128**: 644–667.

Buehner M, Charron M. 2007. Spectral and spatial localization of background-error correlations for data assimilation. *Q. J. R. Meteorol. Soc.* **133**: 615–630.

Corazza M, Kalnay E, Patil DJ, Yang SC, Morss R, Cai M, Szunyogh I, Hunt BR, Yorke JA. 2003. Use of the breeding technique to estimate the structure of the analysis 'errors of the day'. *Nonlinear Processes in Geophys.* **10**: 233–243.

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

Courtier P, Thépaut J-N, Hollingsworth A. 1994. A strategy for operational implementation of 4d-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**: 1367–1387.

Courtier P, Andersson E, Heckley W, Pailleux J, Vasiljevic D, Hollingsworth A, Fisher M, Rabier F. 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc.* **124**: 1783–1807.

Cullen MJP. 2003. Four-dimensional variational data assimilation: A new formulation of the background-error covariance matrix based on a potential-vorticity representation. *Q. J. R. Meteorol. Soc.* **129**: 2777–2796.

Daley R. 1991. *Atmospheric Data Analysis*. Cambridge University Press: Cambridge, UK.

Deckmyn A, Berre L. 2005. A wavelet approach to representing background error covariances in a limited area model. *Mon. Weather Rev.* **133**: 1279–1294.

Dee DP. 2005. Bias and data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3323–3343.

Dee DP, Da Silva AM. 2003. The choice of variable for atmospheric moisture analysis. *Mon. Weather Rev.* **131**: 155–171.

de Pondeca MSFV, Manikin G, Parrish DF, Purser RJ, Wu WS, DiMego G, Derber JC, Benjamin SG, Horel J, Lazarus S, Anderson L, Colman B, Mandt B. 2007. 'The status of the real time mesoscale analysis system at NCEP'. Preprints, 22nd Conf. on WAF/18th Conf. on NWP, Park City, UT. American Meteorological Society, 4A.5.

Derber J, Bouttier F. 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus* **51A**: 195–221.

Derber J, Rosati A. 1989. A global oceanic data assimilation system *J. Phys. Oceanogr.* **19**: 1333–1347.

Desroziers G. 1997. A coordinate change for data assimilation in spherical geometry of frontal structures. *Mon. Weather Rev.* **125**: 3030–3038.

Desroziers G, Lafore JP. 1993. A coordinate transform for objective frontal analysis. *Mon. Weather Rev.* **121**: 1531–1553.

Ehrendorfer M. 2007. A review of issues in ensemble-based Kalman filtering *Meteorol. Z.* **16**: 795–818.

Ehrendorfer M, Bouttier F. 1998. An explicit low resolution extended Kalman filter: Implementation and preliminary experimentation. ECMWF Tech. Memo., **259** (available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK).

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

Fisher M. 1998. Development of a simplified Kalman filter. ECMWF Tech. Memo, **260** (available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK).

Fisher M. 2003. 'Background error covariance modelling'. Pp. 45–64 in ECMWF Seminar on Recent developments in data assimilation for atmosphere and ocean, 8–12 September 2003. ECMWF: Reading UK.

Fisher M. 2004. 'Generalized frames on the sphere, with application to background error covariance modelling'. Pp. 87–102 in ECMWF Seminar on Recent developments in numerical methods for atmosphere and ocean modelling, 6–10 September 2004. ECMWF: Reading UK.

Fisher M. 2007. 'The sensitivity of analysis errors to the specification of background error covariances'. Pp. 27–36 in ECWMF Workshop Proceedings on Flow-dependent aspects of Data Assimilation, 11–13 June 2007. ECMWF: Reading, UK.

Fisher M, Andersson E. 2001. Developments in 4D-Var and Kalman filtering. ECMWF Tech. Memo, **347** (available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK).

Fisher M, Courtier P. 1995. Estimating the covariance matrices of analysis and forecast error in variational data assimilation. ECMWF Tech. Memo, **220** (available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK).

Fischer C, Montmerle T, Berre L, Auger L, Ştefănescu SE. 2005. An overview of the variational assimilation in the ALADIN/France numerical weather prediction system. *Q. J. R. Meteorol. Soc.* **131**: 3477–3492.

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

Gauthier P, Charette C, Fillion L, Koclas P, Laroche S. 1999. Implementation of a 3d variational data assimilation system at the Canadian Meteorological Centre. Part I: The global analysis. *Atmosphere Ocean* **37**: 103–156.

Giering R, Kaminski T. 1998. Recipes for adjoint code construction. *Acm. T. Math. Software* **24**: 437–474.

Gustafsson N, Hörnquist S, Lindskog M, Rantakokko J, Berre L, Navascues B, Huang X, Mogensen K, Thorsteinsson S. 1999. Three-dimensional variational data assimilation for a high resolution limited area model (HIRLAM). HIRLAM Tech. Report **40** (available from HIRLAM, c/o Met Ëireann, Glasnevin Hill, Dublin 9, Ireland).

Gustafsson N, Berre L, Hornquist S, Huang XY, Lindskog M, Navascues B, Mogensen KS, Thornsteinsson S. 2001. Three-dimensional variational data assimilation for a limited area model. *Tellus* **53A**: 425–446.

Haltiner GJ, Williams RT. 1980. *Numerical Prediction and Dynamic Meteorology*, 2nd edition. John Wiley: London.

Hello G, Bouttier F. 2001. Using adjoint sensitivity as a local structure function in variational data assimilation. *Nonlinear Processes in GeoPhys.* **8**: 347–355.

Hollingsworth A, Lönnberg P. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus* **38A**: 111–136.

Hólm E, Andersson E, Beljaars A, Lopez P, Mahfouf JF, Simmons A, Thépaut J-N. 2002. Assimilation and modelling of the hydrological cycle: ECMWF's status and plans. ECMWF Tech. Memo, **383** (available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK).

Honda Y, Nishijima M, Koizumi K, Ohta Y, Tamiya K, Kawabata T, Tsuyuki T. 2005. A pre-operational variational data assimilation system for a non-hydrostatic model at the Japan Meteorological Agency: Formulation and preliminary results. *Q. J. R. Meteorol. Soc.* **131**: 3465–3475.

Hoskins B. 1975. The geostrophic momentum approximation and the semi-geostrophic equations. *J. Atmos. Sci.* **32**: 233–242.

Hoskins B, Bretherton F. 1972. Atmospheric frontogenesis models: Mathematical formulation and solution. *J. Atmos. Sci.* **29**: 11–27.

Houtekamer PL, Mitchell HL. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **129**: 123–137.

Ide K, Coutier P, Ghil M, Lorenc AC. 1997. Unified notation for data assimilation: operational, sequential and variational. *J. Meteorol. Soc. Jpn* **75**: 181–189.

Ingleby NB. 2001. The statistical structure of forecast errors and its representation in the Met Office global three-dimensional variational data assimilation system. *Q. J. R. Meteorol. Soc.* **127**: 209–231.

Isaksen L, Fisher M, Berner J. 2007. 'Use of analysis ensembles in estimating flow-dependent background error variance'. Pp. 65–86 in ECWMF Workshop Proceedings on Flow-dependent aspects of data assimilation, 11–13 June 2007. ECMWF: Reading, UK.

JMA. 2007. Outline of the Operational Numerical Weather Prediction at the Japan Meteorological Agency (available from http://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline-nwp/index.htm).

Johnson C, Hoskins BJ, Nichols NK. 2005. Filtering and interpolation in 4d-Var.: Filtering and interpolation *Q. J. R. Meteorol. Soc.* **131**: 1–19.

DOI: 10.1002/qj

Kalnay E. 2003. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press: Cambridge, UK.

Kalnay E, Toth Z. 1994. 'Removing growing errors in the analysis'. Pp. 212–215 in Preprints, 10th Conf. Numerical Weather Prediction. Portland, OR. American Meteorological Society.

Kalnay E, Anderson DLT, Bennett AF, Busalacchi AJ, Cohn SE, Courtier P, Derber JC, Lorenc AC, Parrish DF, Purser RJ, Sato N, Schlatter T. 1997. Data assimilation in the ocean and atmosphere: What should be next? *J. Meteorol. Soc. Jpn* **75**: 489–496.

Katz D. 2007. *The application of PV-based control variable transformations in variational data assimilation*. PhD Thesis, Reading University, Department of Mathematics.

Lahoz WA, Geer AJ. 2003. 'Some challenges in the assimilation of stratosphere/tropopause satellite data'. Pp. 117–136 in ECMWF/SPARC Workshop on Modelling and Assimilation for the Stratosphere and Tropopause, 23–26 June 2003. ECMWF: Reading UK.

Lahoz WA, Geer AJ, Bekki S, Bormann N, Ceccherini S, Elbern H, Errera Q, Eskes HJ, Fonteyn D, Jackson DR, Khattatov B, Massart S, Peuch V-H, Rharmill S, Ridolfi M, Segers A, Talagrand O, Thornton HE, Vik AF, von Clarmann T. 2006. The assimilation of Envisat Data (ASSET) project. *Atmos. Chem. Phys. Discuss.* **6**: 12769–12824.

Lahoz WA, Fonteyn D, Swinbank R. 2007. Data assimilation of atmospheric constituents: A review. *Atmos. Chem. Phys. Discuss.* **7**: 9561–9633.

Lanczos C. 1988. *Applied Analysis*. Dover Publications: New York.

Laroche S, Gauthier P, St James J, Morneau J. 1999. Implementation of a 3d variational data assimilation system at the Canadian Meteorological Centre. Part II: The regional analysis. *Atmos.–Ocean* **37**: 281–307.

Le Dimet FX, Talagrand O. 1986. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus* **38A**: 97–110.

Lerner JA, Weisz E, Kirchengast G. 2002. Temperature and humidity retrieval from simulated Infrared Atmospheric Sounding Interferometer (IASI) measurements. *J. Geophys. Res.* **107**(D14): ACH 4-1–4-11.

Lorenc AC. 1992. Iterative analysis using covariance functions and filters. *Q. J. R. Meteorol. Soc.* **118**: 569–591.

Lorenc AC. 1997. Development of an operational variational assimilation scheme. *J. Meteorol. Soc. Jpn* **75**: 339–346.

Lorenc AC. 2003a. Modelling of error covariances by 4d-Var assimilation. *Q. J. R. Meteorol. Soc.* **129**: 3167–3182.

Lorenc AC. 2003b. The potential of the ensemble Kalman filter for NWP: A comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**: 3183–3203.

Lorenc AC. 2007. 'Ideas for adding flow-dependence to the Met Office VAR system'. Pp. 1–10 in ECMWF Workshop Proceedings on flow-dependent aspects of data assimilation, 11–13 June 2007. ECMWF: Reading UK.

Lorenc AC, Payne T. 2007. 4D-Var and the butterfly effect: Statistical four-dimensional data assimilation for a wide range of scales. *Q. J. R. Meteorol. Soc.* **133**: 607–614.

Lorenc AC, Ballard SP, Bell RS, Ingleby NB, Andrews PLF, Barker DM, Bray JR, Clayton AM, Dalby T, Li D, Payne TJ, Saunders FW. 2000. The Met Office global three-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **126**: 2991–3012.

Lorenc AC, Roulstone I, White A. 2003. On the choice of control fields in VAR. Forecasting Research Tech. Report **419** (available from Met Office, Fitzroy Road, Exeter, Devon, EX8 3PB, UK).

McIntyre ME. 2003. 'Balanced flow'. Pp. 680–685 in *Encyclopedia of Atmospheric Sciences*, Eds. J.R. Holton, J.A. Curry, J.A. Pyle. Elsevier Science: Amsterdam, the Netherlands.

Ménard R, Chang LP. 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part II: Chi-squared validated results and analysis of variance and correlation dynamics. *Mon. Weather Rev.* **128**: 2672–2686.

Met Office. 1995. Control variable transforms – parameters. VAR Scientific Documentation Paper, **11** (available from Met Office, Fitzroy Road, Exeter, Devon, EX8 3PB, UK).

Obukhov AM. 1954. 'Statistical description of continuous fields', Trudy Geofiz In-ta Akad Nauk SSSR No. 24, 151 3–42 (English translation by Liason Office, Technical Information Centre, Wright-Patterson AFB F-TS–9295/v).

Pagé C, Fillion L, Zwack P. 2007. Diagnosing summertime mesoscale vertical motion: implications for atmospheric data assimilation. *Mon. Weather Rev.* **135**: 2076–2094.

Pannekoucke O, Berre L, Desroziers G. 2007. Filtering properties of wavelets for local background-error correlations. *Q. J. R. Meteorol. Soc.* **133**: 363–379.

Parrish DF, Derber JC. 1992. The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Weather Rev.* **120**: 1747–1763.

Phillips NA. 1986. The spatial statistics of random geostrophic modes and first-guess errors. *Tellus* **38A**: 314–332.

Polavarapu S, Ren S, Rochon Y, Sankey D, Ek N, Koshyk J, Tarasick D. 2005. Data assimilation with the Canadian middle atmosphere model. *Atmos.–Ocean* **43**: 77–100.

Pu ZX, Kalnay E, Parrish D, Wu W, Toth Z. 1997. The use of bred vectors in the NCEP global 3d variational analysis scheme, *Weather and Forecasting* **12**: 689–695.

Purser RJ, Wu WS, Parrish DF, Roberts NM. 2003a. Numerical aspects of the application of recursive filters to variational analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. *Mon. Weather Rev.* **131**: 1524–1535.

Purser RJ, Wu WS, Parrish DF, Roberts NM. 2003b. Numerical aspects of the application of recursive filters to variational analysis. Part II: Spatially inhomogeneous and anisotropic general covariances. *Mon. Weather Rev.* **131**: 1536–1548.

Rabier F, Mahfouf J.-F, Fisher M, Järvinen H, Simmons AJ, Andersson E, Bouttier F, Courtier P, Hamrud M, Haseler J, Hollingsworth A, Isaksen L, Klinker E, Saarinen S, Temperton C, Thépaut J.-N, Undén P, Vasiljevic D. 1997. Recent experimentation on 4D-Var and first results from a Simplified Kalman Filter ECMWF Tech. Memo., **240** (available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK).

Rawlins F, Ballard SP, Bovis KJ, Clayton AM, Li D, Inverarity GW, Lorenc AC, Payne TJ. 2007. The Met Office global four-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **133**: 347–362.

Raymond WH, Garder A. 1991. A review of recursive and implicit filters. *Mon. Weather Rev.* **119**: 477–495.

Riishøjgaard LP. 1998. A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus* **50A**: 42–57.

Sadiki W, Fischer C. 2005. A posteriori validation applied to the 3D-VAR Arpège and Aladin data assimilation systems *Tellus* **57A**: 21–34.

Segers AJ, Eskes HJ, Van der ARJ, Van Oss RF and Van Velthoven PFJ. 2005. Assimilation of GOME ozone profiles and a global chemistry-transport model using a Kalman filter with anisotropic covariance. *Q. J. R. Meteorol. Soc.* **131**: 477–502.

Semple AT. 2002. 'An error breeding system for the Met Office "new dynamics" '. Forecasting Research Tech. Report **413** (available from Met Office, Fitzroy Road, Exeter, Devon, EX8 3PB, UK).

Simmons A. 2003. 'Observations, assimilation and the improvement of global weather prediction – some results from operational forecasting and ERA-40'. Pp. 1–28 in ECMWF Seminar on Recent developments in data assimilation for atmosphere and ocean, 8–12 September 2003. ECMWF: Reading UK.

Skamarock WC, Klemp JB, Dudia J, Gill GO, Barker DM, Duda MG, Huang X-Y, Wang W, Powers JG. 2008. A description of the Advanced Research WRF Version 3. NCAR Technical Note **475** (available from http://www.wrf-model.org).

Stajner I, Riishøjgaard LP, Rood RB. 2001. The GEOS ozone data assimilation system: specification of error statistics. *Q. J. R. Meteorol. Soc.* **127**: 1069–1094.

Sun J, Guo Y, Lim E, Huang X, Xiao Q, Sugimoto S. 2008. 'Assimilation and forecasting experiments using radar observations and WRF 4DVar'. Proc. of ERAD 2008, 5th European Conf. on radar in meteorology and hydrology, 30 June–4 July 2008, Helsinki, Finland. Finnish Meteorological Institute: Helsinki, Finland.

Thépaut J.-N, Hoffman R, Courtier P. 1993. Interaction of dynamics and observations in a four dimensional variational assimilation. *Mon. Weather Rev.* **121**: 3393–3413.

Toth Z, Kalnay E. 1993. Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**: 2317–2330.

Toth Z, Kalnay E. 1997. Ensemble forecasting at NCEP: The breeding method. *Mon. Weather Rev.* **125**: 3297–3318.

Weaver A, Courtier P. 2001. Correlation modelling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc.* **127**: 1815–1846.

Weaver AT, Deltel C, Machu E, Ricci S, Daget N. 2005. A multivariate balance operator for variational ocean data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3605–3626.

Wlasak M, Nichols NK, Roulstone I. 2006. Use of potential vorticity for incremental data assimilation. *Q. J. R. Meteorol. Soc.* **132**: 2867–2886.

Wu WS, Purser RJ, Parrish DF. 2002. Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Weather Rev.* **130**: 2905–2916.

Žagar N, Gustafsson N, Källén E. 2004a. Dynamical response of equatorial waves in four-dimensional variational data assimilation. *Tellus* **56A**: 29–46.

Žagar N, Gustafsson N, Källén E. 2004b. Variational data assimilation in the Tropics: The impact of a background error covariance constraint. *Q. J. R. Meteorol. Soc.* **130**: 103–125.

Zhang F. 2005. Dynamics and structure of mesoscale error covariance of a winter cyclone estimated through short-range ensemble forecasts. *Mon. Weather Rev.* **133**: 2876–2893.

Zupanski D. 1997. A general weak constraint applicable to operational 4d-Var data assimilation systems. *Mon. Weather Rev.* **125**: 2274–2291.

Zupanski M, Zupanski D, Vukicevic T, Eis K, Vonder Haar T. 2005. CIRA/CSU four dimensional variational data assimilation system. *Mon. Weather Rev.* **133**: 829–843.

DOI: 10.1002/qj