# Gravity wave drag estimation from global analyses using variational data assimilation principles. I: Theory and implementation

By M. PULIDO* and J. THUBURN

*Department of Meteorology, University of Reading, UK*

(Received 2 August 2004; revised 2 February 2005)

## SUMMARY

A novel technique to estimate gravity wave drag from global-scale analyses is presented. It is based on the principles of four-dimensional variational data assimilation, using a dynamical model of the middle atmosphere and its adjoint. The global analyses are treated as observations. A cost function that measures the mismatch between the model state and observations is defined. The control variables are the components of the three-dimensional gravity wave drag field, so that minimization of the cost function gives the optimal gravity wave drag field. The minimization is performed using a conjugate gradient method, with the adjoint model used to calculate the gradient of the cost function.

In this work, we present the theory behind the new technique and evaluate extensively the ability of the technique to estimate the gravity wave drag using so-called twin experiments, in which the 'observations' are given by the evolution of the dynamical model with a prescribed gravity wave drag. The results show that the technique can estimate accurately the prescribed gravity wave drag. When the cost function is suitably defined, there is good convergence of the minimization scheme under realistic atmospheric conditions. We also show that the cost function gradient is well approximated taking into account only adiabatic processes. We note some limitations of the technique for estimating gravity wave drag in tropical regions if satellite temperature measurements are the only observational information available.

KEYWORDS: Adjoint model Middle atmosphere Parameter estimation Twin experiments

## 1. INTRODUCTION

Small-scale gravity waves have profound influence on the general circulation of the middle atmosphere. They propagate upwards from the troposphere, and give rise to a convergence of pseudo-momentum flux, or 'drag', where they break or dissipate in the middle atmosphere. The drag is responsible for closing the mesospheric jets and for the reversal of the mesospheric meridional temperature gradient (e.g. Lindzen 1981); it also affects the lower stratosphere and upper troposphere (e.g. Palmer *et al.* 1986), and is thought to be a significant component of the driving of the quasi-biennial oscillation (e.g. Lindzen and Holton 1968; Baldwin *et al.* 2001).

The drag due to gravity waves cannot be measured directly from observations. However, since the waves produce notable effects in the general circulation, by using large-scale observations and an inverse method it is possible to infer the gravity wave drag (GWD) that is producing those effects. This concept has been used by budget studies, where the residual term in the mean momentum equations is attributed to the GWD field.

To date, budget studies have been based on either a zonal mean (e.g. Hamilton 1983; Shine 1989; Marks 1989; Alexander and Rosenlof 2003) or a time mean (Klinker and Sardeshmukh 1992) of the momentum equations. Thus, they suffer from the limitation of only giving information on the zonal mean or the time mean of the GWD field, and in some cases only the zonal component.

A particular motivation for quantifying GWD is the need to represent it in general circulation models. Small-scale gravity waves cannot be resolved in current general-circulation models; instead the GWD is taken into account by means of parametrizations (e.g. Hines 1997; Warner and McIntyre 1996). These parametrizations have improved

* Corresponding author, present affiliation: Department of Physics, FACENA, Northeastern National University, Av. Libertad 5460, (3400) Corrientes, Argentina. e-mail: pulido@exa.unne.edu.ar

the results of general-circulation models (e.g. Scaife *et al.* 2002). However, they contain many simplifying assumptions and tunable parameters. These parameters are chosen subjectively to give a good representation of the phenomenon that is being modelled. Given the arbitrariness of this tuning, a set of parameters that allow, for instance, the characteristics of the quasi-biennial oscillation to be reproduced, may give unrealistic features in high latitudes. Therefore, there is a current need for observational evidence on the GWD to help improve parametrization schemes.

There have been some attempts to constrain the parameters using the estimated GWD from budget studies, but they cannot determine all the parameters since there are more unknown parameters in the drag schemes than information on the drag field (Alexander and Rosenlof 2003). Observational knowledge of meridional and zonal components of the GWD, and their distribution in space and time, can be a key point in overcoming this difficulty.

In this work, we present a novel technique that uses data assimilation principles to estimate GWD from observations. The technique is based on four-dimensional variational assimilation (4D-Var) which allows the representation of the flow evolution from the initial conditions to the final state keeping information of the whole trajectory. As a result, in principle, the zonal and meridional components of the GWD field in three-dimensional space can be determined, including its evolution with time. The estimated field is the GWD that best reproduces the observed large-scale fields during the model evolution from the initial time to the final state. In this way, the present technique is able not only to determine the three-dimensional field of the meridional and zonal components of the GWD but also avoids another of the problems related to budget studies: it is able to distinguish between the local and remote GWD that produce perturbations contributing to the same point at a given time. We apply the technique to estimating middle atmosphere GWD, since GWD is more significant in the middle atmosphere than in the troposphere and the implementation is simpler since it does not require the use of a full tropospheric general-circulation model.

The paper is organized as follows. The next section presents the theoretical details of the technique. Section 3 explains the practical implementation including a description of the dynamical model and its adjoint. Then, we present a series of simulations to show the technique performance, for ideal and realistic flow conditions. Finally, we examine how the accuracy of the method could be affected by errors in the model radiation scheme.

## 2. Theoretical background

Variational data assimilation offers an objective way of estimating the initial conditions or other unknown parameters of a numerical model; for an introduction see Errico (1997). In general, a cost function or 'model error' that measures the mismatch between the observations and the state of the model is defined. In turn, the model state is a function of the unknown parameters. The cost function is minimized using the unknown parameters as control variables, and the resulting optimal values of the control variables give the best estimate of the unknown parameters.

In this study we use a three-dimensional, time-dependent model of the middle atmosphere, described in section 3(a). The control variables are the components, $\mathbf{X}$, of the GWD field. As a first approximation, we assume the GWD is independent of time within the assimilation window. The field is specified directly on the same grid as the model state; therefore the number of dimensions of the control space is $2N$ where $N$ is the number of model grid points.

We define the cost function by

$$J = \frac{1}{2}\sum_{i=1}^{n}(H[\mathbf{y}_i] - \mathbf{x}_i)^{\mathrm{T}}\mathbf{R}^{-1}(H[\mathbf{y}_i] - \mathbf{x}_i), \tag{1}$$

where $\mathbf{x}_i$ is the model state at time $t_i$, $H$ is the operator that transforms the observed variables, $\mathbf{y}_i$, to the model space and $\mathbf{R}$ is a certain matrix, discussed below. The state $\mathbf{x}_i$ is given by the model evolution from $t_0$ to $t_i$

$$\mathbf{x}_i = M(\mathbf{x}_0, \mathbf{X}, t_i), \tag{2}$$

where $\mathbf{x}_0$ is the known initial condition.

Note the model variable $\mathbf{x}_i$ is not necessarily the same as the observed variable $\mathbf{y}_i$, since $H$ may represent not only a grid interpolation but also a variable transformation. For instance, wind can be represented by the velocity field, $(u, v)$, or by the absolute vorticity and divergence, $(f + \zeta, \delta)$. Under the hydrostatic balance approximation, temperatures may be transformed to geopotential height or to pseudo-density.

One difference in the cost function definition (1) from the usual form used in 4D-Var is that observations are transformed to the model grid and variables rather than the other way round. This is possible in our case because it does not involve any complicated inverse modelling (unlike retrieving temperatures from radiances, for example) and it saves computing time since the observations only need to be transformed once; the estimated GWD is not significantly affected.

Another difference in the cost function from the usual 4D-Var form is the omission of a term of the form $(\mathbf{X} - \mathbf{X}_{\mathrm{b}})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{X} - \mathbf{X}_{\mathrm{b}})$ measuring the difference between the drag $\mathbf{X}$ and some background or a priori estimate of the drag, $\mathbf{X}_{\mathrm{b}}$. This omission is equivalent to the assumption that we have a priori perfect ignorance of the GWD (a reasonable first approximation), so that the background error covariance $\mathbf{B}$ is infinite. In the usual form of 4D-Var, a background term is essential to make the problem well posed, because there are far fewer pieces of observational information than degrees of freedom in the control variables. Because in our problem the observations are three-dimensional global analyses of wind and temperature, the problem is well posed without the background term. Future work could also include such a background term using a GWD parametrization, or a climatology produced using the current method, to give $\mathbf{X}_{\mathrm{b}}$. However, the estimation of $\mathbf{B}$ would be very difficult.

We assume that our ignorance of the GWD is much greater than the uncertainty in the observations, and therefore in the cost function (1) we make no explicit allowance for errors in the observed data, which is used to define the model initial conditions as well as the $\mathbf{y}_i$. By a suitable choice of the cost function and control variables, we can make some allowance for the fact that the observed rotational flow is likely to be more accurate than the observed divergent flow (see section 4).

In the usual 4D-Var, $\mathbf{R}$ must be the observation error covariance matrix, so that minimization of $J$ gives the optimal balance between observational and background errors. In the absence of the background term, there is considerable freedom to choose different $\mathbf{R}$s without affecting the $\mathbf{X}$ that minimizes the cost function. We exploit this freedom to choose $\mathbf{R}$ to give good conditioning, and hence fast convergence, of the minimization algorithm.

Since the initial state is prescribed, the only unknown quantity in (1) and (2) is the drag vector, $\mathbf{X}$. Therefore, the minimum of the cost function will determine the optimal drag. The model evolution from $t = t_0$ to $t_n$ with the given initial conditions and the optimal drag minimizes the mismatch between the observations and model state along the entire temporal window from $t = t_0$ to $t_n$. In practice, we have used $n = 1$, i.e. a single observation time at the end of the assimilation window.

We use a conjugate gradient minimization algorithm which requires at each iteration the gradient of the cost function with respect to the drag. In practice, the gradient of the cost function is calculated by means of an adjoint model (see section 3(b)).

The rate of convergence depends on the condition number of a certain matrix. Let $\widehat{\mathbf{M}}_i$ be the tangent linear model corresponding to the model $M$ defined in (2), linearized about the control case, $\mathbf{x}_\mathrm{c}$, i.e. the evolution of the model with zero drag. Then

$$\mathbf{x}'_i = \mathbf{x}_i - \mathbf{x}_{\mathrm{c}i} \approx \widehat{\mathbf{M}}_i \mathbf{X}. \tag{3}$$

Representing the true drag by $\mathbf{X}^*$, then, if the model is perfect,

$$\mathbf{x}'^*_i = H[\mathbf{y}_i] - \mathbf{x}_{\mathrm{c}i} \approx \widehat{\mathbf{M}}_i \mathbf{X}^*. \tag{4}$$

Substituting in (1) gives

$$J \approx \frac{1}{2}\sum_{i=1}^{n} (\mathbf{X} - \mathbf{X}^*)^\mathrm{T} \widehat{\mathbf{M}}_\mathbf{i}^\mathrm{T} \mathbf{R}^{-1} \widehat{\mathbf{M}}_\mathbf{i} (\mathbf{X} - \mathbf{X}^*). \tag{5}$$

It is the condition number of the matrix

$$\sum_{i=1}^{n} \widehat{\mathbf{M}}_i^\mathrm{T} \mathbf{R}^{-1} \widehat{\mathbf{M}}_i$$

that determines the rate of convergence; a condition number close to 1 gives fast convergence while a large condition number gives slow convergence. The choice of which $\mathbf{R}^{-1}$ we use in practice is discussed in sections 3(c) and 4(c).

Like 4D-Var, this drag estimation technique relies on the assumption that the dependence of the model state $\mathbf{x}_i$ on the control variables $\mathbf{X}$ is approximately linear. If it is exactly linear, then the cost function is exactly quadratic and the minimization algorithm converges as theoretically predicted. If the dependence is nonlinear, then the cost function will not be exactly quadratic and convergence may be impaired. When the nonlinearity is very strong, the cost function may have multiple minima, and the minimization algorithm could converge to the wrong one. We will show later that, for realistic flow field and drag amplitudes and for an assimilation window of one day, linearity is in fact an excellent approximation.

## 3. Technique implementation

In this section we discuss the implementation of the technique. We apply the theoretical ideas developed in the previous section to a specific middle-atmosphere dynamical model. In particular, the components of the developed system, called ASDE (Assimilation System for Drag Estimation), are described.

### (*a*) *Dynamical model*

The dynamical model used in this study models the middle atmosphere from approximately 100 mb to approximately 0.01 mb. It is based on the fully nonlinear, hydrostatic primitive equations, with an isentropic vertical coordinate and a hexagonal-icosahedral horizontal grid (Gregory 1999).

The hydrostatic primitive equations in isentropic coordinates represented in the model are

$$\partial_t \sigma + \nabla \cdot (\sigma \mathbf{u}) + \partial_\theta (\sigma \dot{\theta}) = 0, \tag{6}$$

$$\partial_t (\sigma Q) + \nabla \cdot (\sigma Q \mathbf{u} - \widehat{\mathbf{k}} \times \dot{\theta} \partial_\theta \mathbf{u}) = X_\zeta, \tag{7}$$

$$\partial_t \delta + \nabla \cdot \left[ \sigma Q \widehat{\mathbf{k}} \times \mathbf{u} + \nabla \left( \Psi + \frac{\mathbf{u}^2}{2} \right) + \dot{\theta} \partial_\theta \mathbf{u} \right] = X_\delta, \tag{8}$$

where the model variables are potential vorticity, $Q$, divergence, $\delta$, and pseudo-density, $\sigma \equiv \rho \partial_\theta z$, and $\nabla$ is the horizontal gradient operator. The drag terms are defined by $X_\zeta = \widehat{\mathbf{k}} \cdot \nabla \times \mathbf{X}$ and $X_\delta = \nabla \cdot \mathbf{X}$. The relation between pseudo-density and the Montgomery potential $\Psi$ is given by the expressions

$$\partial_\theta p = -g\sigma \tag{9}$$

$$\partial_\theta \Psi = c_p \left( \frac{p}{p_0} \right)^\kappa, \tag{10}$$

where $p$ is pressure, $p_0 = 10^5$ Pa and $\kappa = R/c_p$.

The model has $p = 0$ at the top, and a bottom boundary condition near the tropopause at a potential temperature of $\theta = 414$ K, where a time-dependent observational Montgomery potential is imposed.

The vertical velocity across isentropes is given by

$$\dot{\theta} = \Phi(\sigma). \tag{11}$$

A parametrization of the radiative transfer $\Phi(\sigma)$ (Shine 1987; Shine and Rickaby 1989) is used to determine $\dot{\theta}$. The scheme includes the radiative effects of $CO_2$, $O_3$ and $H_2O$. The ozone distribution used for the radiation calculation is prescribed using monthly means from a zonally averaged climatology.

The hexagonal-icosahedral horizontal grid used in this study has 2562 cells that corresponds to a horizontal resolution of 480 km. There are 16 vertical levels which lead to a vertical resolution of about 3 km.

### (*b*) *Adjoint model*

The gradient of the cost function is calculated by integrating the dynamical model forwards over the assimilation window, then integrating the adjoint of the tangent linear model, linearized about the forward trajectory, backwards over the assimilation window.

To develop the adjoint model, it is convenient to treat the GWD components as additional state variables satisfying

$$\partial_t \mathbf{X} = \mathbf{0}. \tag{12}$$

The control variables are then the initial values of $\mathbf{X}$, so that this parameter estimation problem can be treated in exactly the same way as an initial value estimation problem.

Part of the code was developed using the Tangent and Adjoint Module Compiler (Giering and Kaminski 1998), but some manual intervention was necessary to obtain efficient codes. The complete forward trajectory is stored in order to evaluate the adjoint matrix each model time step.

As we are concerned with time-scales of the order of a day, the effects of radiative processes are not taken into account in the adjoint calculation, i.e. the sensitivity of $\dot{\theta}$ is neglected. We discuss this point further in section 5.

Preliminary tests during the adjoint development showed that the original flux-limited advection schemes of the dynamical model (Thuburn 1996) could impair the smoothness of the cost function, and hence the convergence of the minimization, due to the artificial nonlinearities introduced by the flux limiter, which was switched on/off many times during the evolution (Thuburn and Haine 2001; Vukićević *et al.* 2001). Because of this, the flux limiter was removed and the linear version of the advection scheme and its adjoint were used in this work.

### (*c*) *Assimilation details*

The minimization is performed by an iterative method; the conjugate gradient method is used to find the next minimization direction, and in each direction the secant method is used to estimate the one-dimensional minimum. The conjugate gradient method offers a good balance between convergence rates and computer memory requirements for problems with a large number of degrees of freedom (Navon and Legler 1987). Newton methods have a quicker rate of convergence (quadratic) but they require storage of the Hessian matrix.

As each minimization iteration is computationally very expensive and the number of degrees of freedom in the control space is large, an efficient minimization algorithm is needed. One of the aims of this study is to show that only a few iterations are needed to achieve an accurate forcing estimation.

The importance of choosing $\mathbf{R}^{-1}$ so that the problem is well conditioned was noted in section 2. The analysis in the Appendix gives information on the form that $\widehat{\mathbf{M}}$ takes, and hence helps to choose a suitable $\mathbf{R}^{-1}$. On very short times, $t \ll \omega^{-1}$ (see (A.5)), and with our choice of control variables, $\widehat{\mathbf{M}}$ takes a particularly simple diagonal form; it follows that the cost function

$$J = \frac{1}{2}\sum_{i=0}^{n}\sum_{k=1}^{N}(\delta_{ik} - \delta_{ik}^{*})^2 + \{\overline{\sigma}(\theta)\}^2(Q_{ik} - Q_{ik}^{*})^2 \tag{13}$$

leads to a perfectly conditioned problem under idealized conditions. Here $i$ is the time index, $k$ is the index on the three-dimensional grid, and $\overline{\sigma}$ is the horizontally averaged pseudo-density. On longer time-scales, under more realistic conditions, the analysis suggests that the following form, although not perfectly conditioned, should be a good choice for the relative weights of the three terms and the relative weights of the different vertical levels:

$$J = \frac{1}{2}\sum_{i=0}^{n}\sum_{k=1}^{N}(\delta_{ik} - \delta_{ik}^{*})^2 + \{\overline{\sigma}(\theta)\}^2(Q_{ik} - Q_{ik}^{*})^2 + \{\tau\overline{\sigma}(\theta)\}^{-2}(\sigma_{ik} - \sigma_{ik}^{*})^2. \tag{14}$$

Here $\tau$ is a tunable time-scale; experimentation in realistic conditions showed that a value $\tau = 4 \times 10^4$ s worked well.

As control variables, we use the curl $X_\zeta$ and divergence $X_\delta$ of the GWD. This choice is particularly easy to implement in our dynamical model. More importantly, it provides a separation between the dynamically important $X_\zeta$, which forces a geostrophic growing response in $\sigma$ and $Q$ on long time-scales, and the less important $X_\delta$, whose response is a steady $\zeta$ and $\sigma$ anomaly and forced gravity waves. These fields $X_\zeta$ and $X_\delta$ are specified on the same grid as model prognostic variables. The number of degrees of freedom in the control space for the standard setting of ASDE is of the order of $10^5$. Tests were performed in order to assess the dependence of the estimation on the number of degrees of freedom. Specifically, the components of the drag were expressed as truncated series of spherical harmonics and only a limited number of modes were retained to reduce the number of degrees of freedom and risk of noise. The results showed that, with the same number of minimization iterations, there were no significant differences between the large-scale GWD estimated with the full control space and with the reduced control space. Indeed the estimated drag with the full control space was not noisy, and it was computationally cheaper since the transformations between grid and spherical harmonics are avoided.

For the tests reported below, the assimilation window is taken to be $\tau = 24$ h and the final time is the only observation time. The chosen assimilation window length is a compromise between frequency of the available analyses which are taken as the observations and computer resources needed to store the whole forward trajectory. But, it is also a reasonable scale for GWD variability.

If we want to estimate the GWD for a period longer than a single assimilation window, the GWD estimations are performed for each successive 24 h assimilation window. For the first window, the initial conditions are taken from the analyses. For subsequent windows the initial conditions are taken as the final model state from the previous assimilation window when the model is run with the best estimate of the GWD; the optimization ensures that these initial conditions are close to the observations. This procedure has the advantage that the model state evolves continuously between successive assimilation windows, rather than being repeatedly re-initialized. In particular, the model remains balanced and has a closed angular momentum budget. A potential disadvantage is that the model could experience a drift (for example in horizontal mean $\sigma$ due to radiation errors) which cannot be corrected by the estimated drag. To avoid such problems we have used this procedure for up to one month at a time, then we re-initialized the model from observations.

## 4. Idealized twin experiments

In the following sections we describe tests carried out to demonstrate that the method works, to investigate the validity of some of the approximations and assumptions made, and to tune some of the choices in the method such as the exact form of the cost function and the number of conjugate gradient iterations used.

To test the technique we use 'twin experiments', in which the same dynamical model with a prescribed drag is used to calculate the 'observation', and then the ASDE is applied to the 'observation' in order to see how well the prescribed drag can be estimated. For these tests, the model is run for one day from known initial conditions and with the true GWD specified. The state at 24 h is then taken as the observation.

### (*a*) *Estimation of a prescribed analytical GWD*

The prescribed GWD is a three-dimensional field with both zonal and meridional components. It is defined by a three-dimensional Gaussian centred at a latitude of $-45°$, as shown in Fig. 1. The maximum zonal GWD is 15 m s$^{-1}$day$^{-1}$ and the maximum meridional GWD is 7.5 m s$^{-1}$day$^{-1}$. Note the chosen intensities are similar to the zonal intensities inferred from observations using budget studies (Hamilton 1983; Shine 1989).

The evolution of the model is adiabatic, computed starting from rest with an isothermal atmosphere, $T = 250$ K, and constant Montgomery potential at the bottom boundary. From this evolution, we take the model state at $t$ =24 h as the observations which are also shown in Fig. 1. The flow responds to the drag by producing a westward

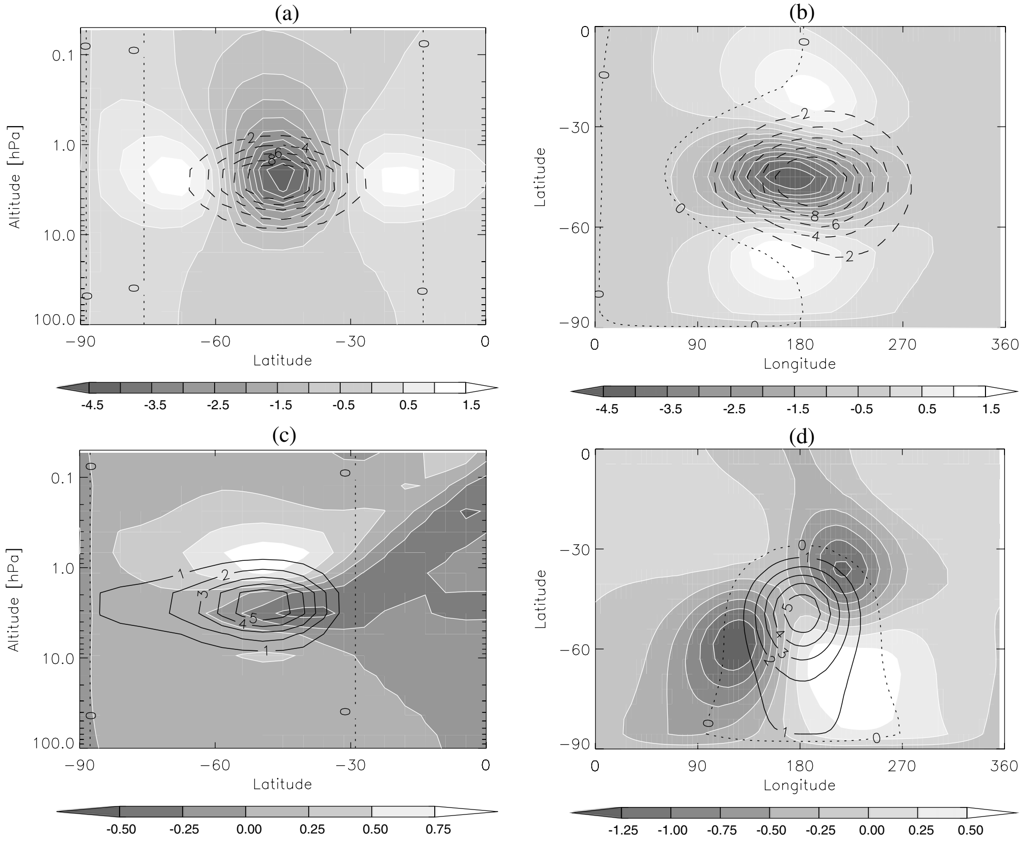

Figure 1. Prescribed zonal component of gravity wave drag (m s$^{-1}$day$^{-1}$, black contours, solid denoting positive and dashed negative), and $u$ perturbations (m s$^{-1}$, white contours and shading) at $t = 24$ h resulting from the evolution of the model with the prescribed drag; (a) is a vertical section at 180° longitude, and (b) a horizontal section at 1.9 hPa. (c) and (d) are as (a) and (b), respectively, but showing meridional components and $v$ perturbations. Transformation to the model grid and variables distorts the drag field slightly; the fields shown are those actually 'felt' by the model.

jet located at the centre of the drag with two eastward jets at the latitudinal extremes and a meridional circulation in the height–latitude plane.

Then the ASDE is applied to the observations in order to estimate the prescribed drag. The first-guess drag is set to zero (we assume there is no previous information). In this first experiment we assume that the three state variables, $\sigma$, $Q$, $\delta$, are observed and therefore the cost function was defined by (14) with $\tau = 4 \times 10^4$ s.

The convergence of the technique is very good; it was found that 25 minimization iterations are enough to achieve an accuracy of the order of 1 m s$^{-1}$day$^{-1}$ in the drag estimation. Figure 2 shows the results of the experiment. There is a good agreement between the estimated and prescribed drag. The intensity and shape of the prescribed drag are well estimated. The largest errors are in the meridional component, whose maximum intensity is underestimated by 1.2 m s$^{-1}$day$^{-1}$, and its pattern is elongated towards the South Pole.

The variational method estimate of the drag can be compared with a crude budget-based estimate given by $\mathbf{X} \approx \{\mathbf{u}(1\text{ day}) - \mathbf{u}(0)\}/1$ day. For this idealized test case, the budget estimate is given by the shading in Fig. 1, reinterpreted as a drag in m s$^{-1}$day$^{-1}$.

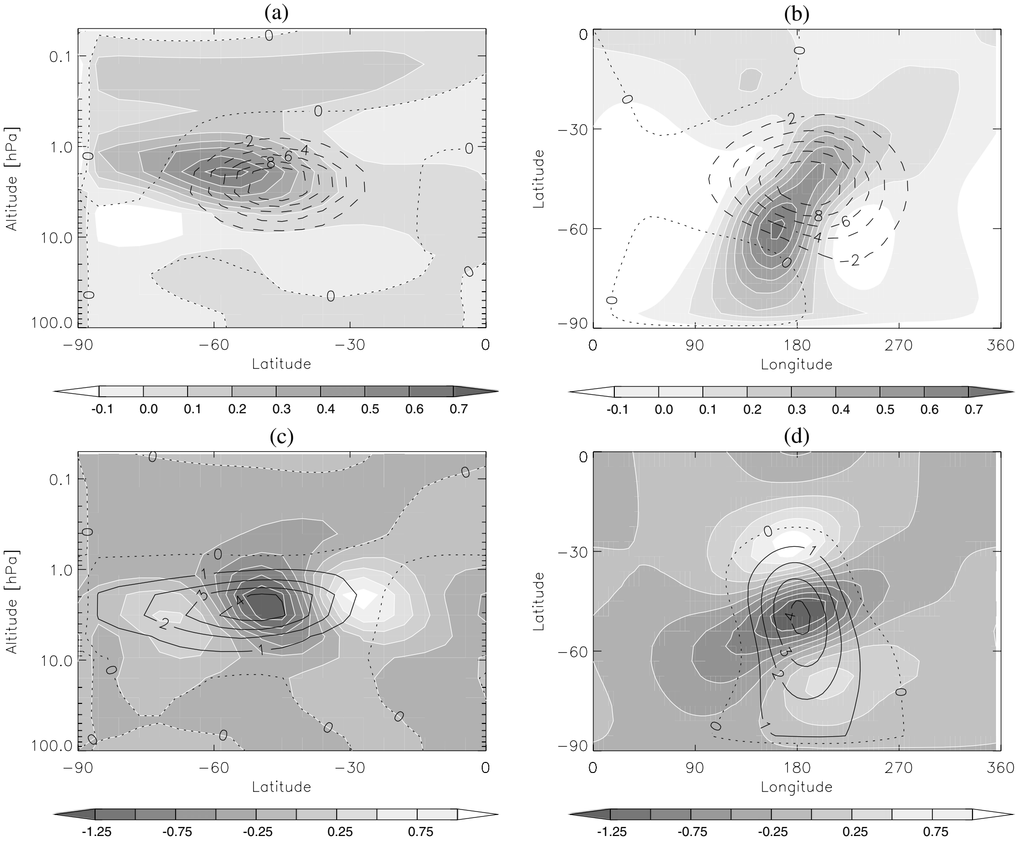

Figure 2. Estimated (a, b) zonal and (c, d) meridional components of gravity wave drag (GWD, black contours) and the errors in the GWD estimation (white contours and shading), both m $\mathrm{s}^{-1}\mathrm{day}^{-1}$. Sections are as Fig. 1.

There are important differences between the budget estimate and the true drag, not only in the amplitude but also in the morphology. On the other hand, the adjoint model is able to trace back these effects and to identify the source of the forcing that produces them (cf. Figs. 2 and 1).

### (*b*) *Convergence of the technique*

As already mentioned, the convergence could be affected by two factors. The first is nonlinearity, which may be produced by the dynamics themselves, especially with long assimilation intervals or large drag values, or may be generated artificially by numerical schemes. These nonlinearities may produce departures of $J$ from the quadratic form or even multiple minima. To assess the influence of nonlinearities, we investigated the geometry of the cost function along the minimization path. Figure 3 shows the cost function and its derivative in a typical minimization direction. The cost function was found to be convex in all the directions. Indeed the derivative of the cost function is extremely close to linear.

Figure 3 also shows a comparison between the derivative of the cost function calculated with the adjoint model and that calculated directly by finite differences from the cost function. The perfect agreement confirms that the adjoint code is calculating the gradient correctly.

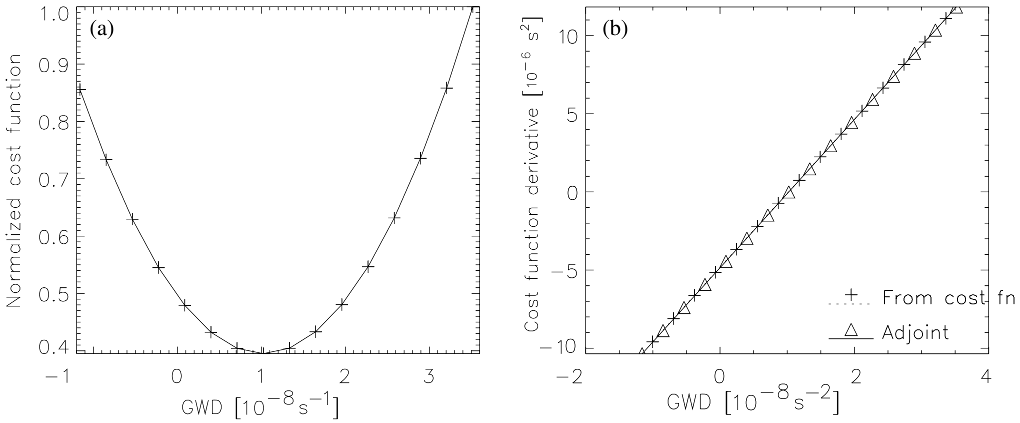

Figure 3. Cost function shape at the 10th minimization direction: (a) cost function and (b) derivative of the cost function, calculated with the adjoint model (solid line) and directly from cost function (dotted line). (There are no visible differences between the curves). Since the control variables are the curl and divergence of the drag, their units are s$^{-2}$.

The second factor that could affect convergence is an anisotropic cost function that results in a poor conditioning. To investigate the convergence rate, Fig. 4(a) shows the cost function and root-mean-square errors in the drag components (all normalized to one initially) versus iteration number. The method is smoothly converging in both the error in the observational variables and in the control variables. Although the control space has dimension of order $10^5$, the cost function decreases by a factor more than $10^2$ in just 25 iterations, showing that, with the choice of control variables and **R** (see (14)), the problem is well conditioned.

For any given number of iterations, the zonal drag $X_x$ is better estimated than the meridional drag $X_y$ (Fig. 4(a)). This appears to be because the curl of the drag $X_\zeta$ is better estimated than the divergence of the drag $X_\delta$ (Fig. 4(b)); note that the zonal mean of $X_x$ is determined entirely by $X_\zeta$, while the zonal mean of $X_y$ is determined entirely by $X_\delta$. The reason why $X_\zeta$ is better estimated than $X_\delta$ can be found in the perturbations that these forcings produce: $X_\zeta$ generates gravity waves and a geostrophic mode, which changes the mean state, while $X_\delta$ generates gravity waves and steady $\sigma$ and $\zeta$ anomalies (see Appendix). The geostrophic mode keeps a simple, local relationship between $X_\zeta$ and $Q$. However, gravity waves are propagating away and then the relationships between $X_\delta$ and $\sigma$ become non-local and scale-dependent. This fact affects the conditioning in the problem of determining $X_\delta$. Besides, some of the generated gravity waves are dissipated so that the information they contain is lost, and the $X_\delta$ forcing them cannot be recovered by the backwards integration.

### (*c*) *Limited observational information*

We have shown in the last section that in ideal conditions there is good convergence of the method towards the true drag. Now we will study some issues related with incomplete observational information. Particularly, we are concerned with how much drag information can be obtained with the ASDE if only one or two observational variables are available. The motivation for examining this is that the main observational input into middle atmosphere analyses is satellite observations of temperature, so that

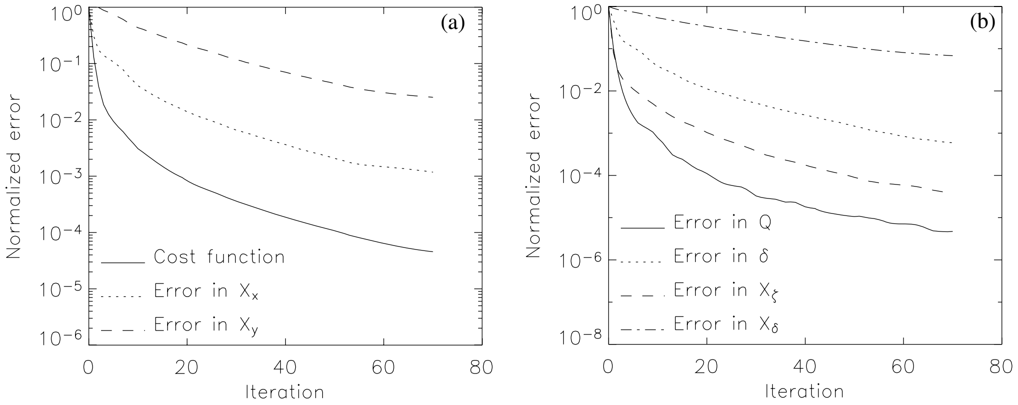

Figure 4. Error as a function of minimization iteration in (a) normalized cost function (solid), and zonal (dotted) and meridional (dashed) drag estimation, and in (b) $Q$ (solid), $\delta$ (dotted), $X_\zeta$ (dashed) and $X_\delta$ (dash-dotted). $Q$ errors are weighted with a $\overline{\sigma}$ factor as in Eq. (14). See text for definitions.

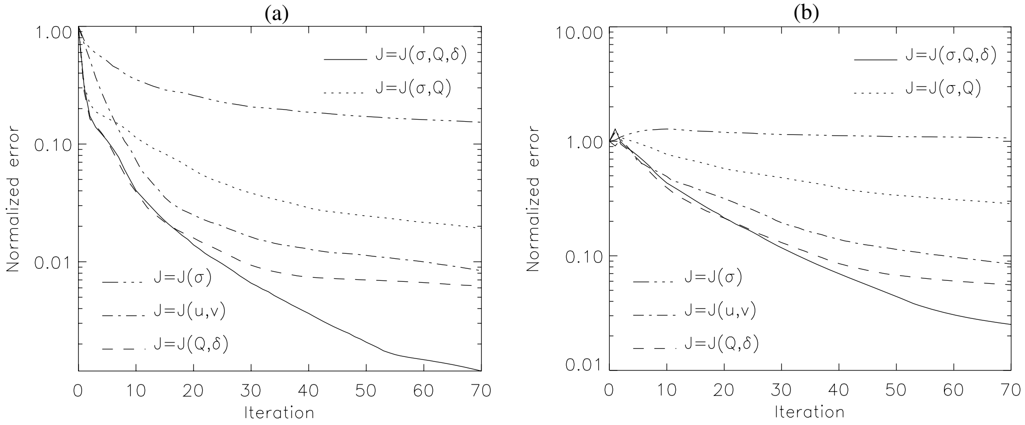

Figure 5. Error as a function of minimization iteration in the estimation of (a) zonal drag and (b) meridional drag, using the methods described in the text.

the unbalanced divergent component of the analysed winds is likely to be less accurate than the balanced rotational component.

The experiment of section 4(a) was repeated using the following alternative cost functions: $J(Q, \delta)$ given by (13), $J(Q, \sigma)$ given by (14) but without the $\delta$ term, and $J(u, v)$ given by

$$J(u, v) = \frac{1}{2}\sum_{k=1}^{N}\{(u_k - u_k^*)^2 + (v_k - v_k^*)^2\}. \tag{15}$$

In all cases the true isothermal resting initial condition is used. Figure 5 shows the convergence of the root-mean-square errors in the estimated $X_x$ and $X_y$, compared with the full case $J(\sigma, Q, \delta)$ given by (14). None of the reduced cost functions perform as well as the full cost function. Nevertheless, they all do converge towards the same solution. For the first 10 to 20 iterations, $J(Q, \delta)$ performs as well as $J(\sigma, Q, \delta)$, but converges much more slowly after that. Finally, although $J(\sigma, Q)$ converges slower than the alternatives tested, it does converge, and has the advantage of not using divergence data, which might be unreliable in practical applications of the method.

Since practical middle atmosphere analyses are dominated by satellite observations of temperature, with wind information coming via the assumption of thermal wind balance, we performed two experiments with the cost function $J(\sigma)$ given by the last term in (14) so that it contains only temperature information and no direct wind information.

The first experiment was the same as that in section 4(a) except for the choice of the cost function. The results are shown in the left panels in Fig. 6. Both the zonal drag and the zonal wind itself are well estimated, particularly in the zonal mean. However the overall errors are worse than the other alternative cost functions (Fig. 5). This reflects the fact that the rotational component of GWD is well estimated while the divergent component is not converging. For this cost function $J(\sigma)$, we are using only one observational variable at only one time and therefore the control space has more degrees of freedom than the observational space. Therefore we should not expect to be able to determine both components of GWD.

The second experiment was similar to the first except the prescribed 'true' drag was centred on the equator. The results are shown in the right panels of Fig. 6. In this case the assimilation procedure does not converge to the correct values of either component of the GWD, or the wind (see Fig. 6(f)).

In theory, the perturbation to the control state produced by a drag component, say $X_\zeta$, will generate a perturbation in $Q$ which will induce a perturbation in the other state variables, $\delta$ and $\sigma$. Therefore, the drag $X_\zeta$ may be determined using observations of, for instance, $\sigma$ alone at several times within the assimilation window. However, the linear dynamical equations are decoupled near the equator, where $f \to 0$. In this simplified case, a perturbation in $Q$ may not induce a related perturbation in $\sigma$ at all, as seen in the solution (A.7)–(A.9), and so we cannot obtain information on the drag by observing $\sigma$.

## 5. Twin experiments in realistic conditions

### (*a*) *Real initial conditions*

So far, the study was kept as simple as possible to focus on the theoretical points of the technique. In this section we will assess what happens under realistic conditions. The first set of experiments deals with realistic initial conditions and realistic bottom-boundary forcing at $\theta_{\mathrm{B}} = 414$ K, both taken from Met Office/Upper Atmosphere Research Satellite analyses (Swinbank and O'Neill 1994). Therefore the model is evolving as a realistic simulation except that radiative heating is still zero, and we have added the same prescribed forcing as in the idealized experiments. Note in this case that the simulations will have a variety of waves produced by the boundary forcing and internally generated motions that are propagating upwards together with the perturbations forced by the GWD, which may interact nonlinearly.

In order to examine the technique in strong wind and shear conditions, we have chosen the initial condition at 1 July 2002 where a very strong jet and also high planetary wave activity are present at the altitude of the GWD (see contours in Fig. 7). In these conditions there are interactions between the response to the GWD and the evolution of the control case. Indeed, the response of the system to the GWD has changed radically compared to the case starting from rest. Figure 7 shows the difference of the observed wind and the control wind (evolution without drag) at 24 h. Both the intensity and the pattern have changed because of interactions between the control flow and the forcing response. (Compare the shading in Fig. 7 to the shading in Fig. 1.)

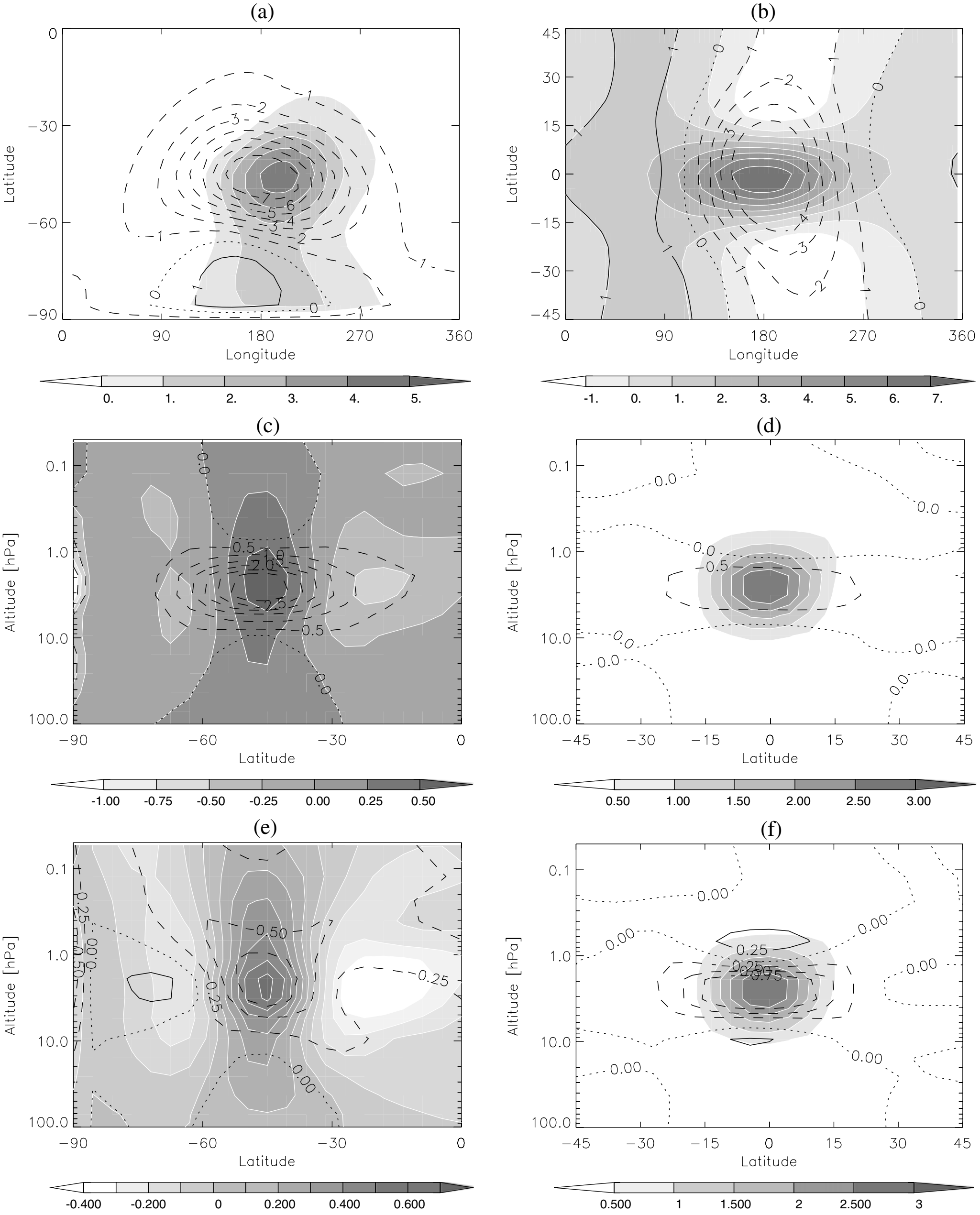

Figure 6. Estimation of gravity wave drag (GWD) using pseudo-density as the observed variable. Black contours (dashed values negative) are estimated fields and white contours with shading are the errors in the estimation (both m $s^{-1}$day$^{-1}$). (a) and (b) show estimated zonal GWD and its errors at 1.9 hPa, (c) and (d) show estimated zonal mean zonal GWD and its errors at 180° longitude, and (e) and (f) show zonal mean zonal wind perturbation and its errors at 180° longitude. (a), (c) and (e) show the case with GWD centred at −45°, and (b), (d) and (f) the case with GWD centred at the equator.

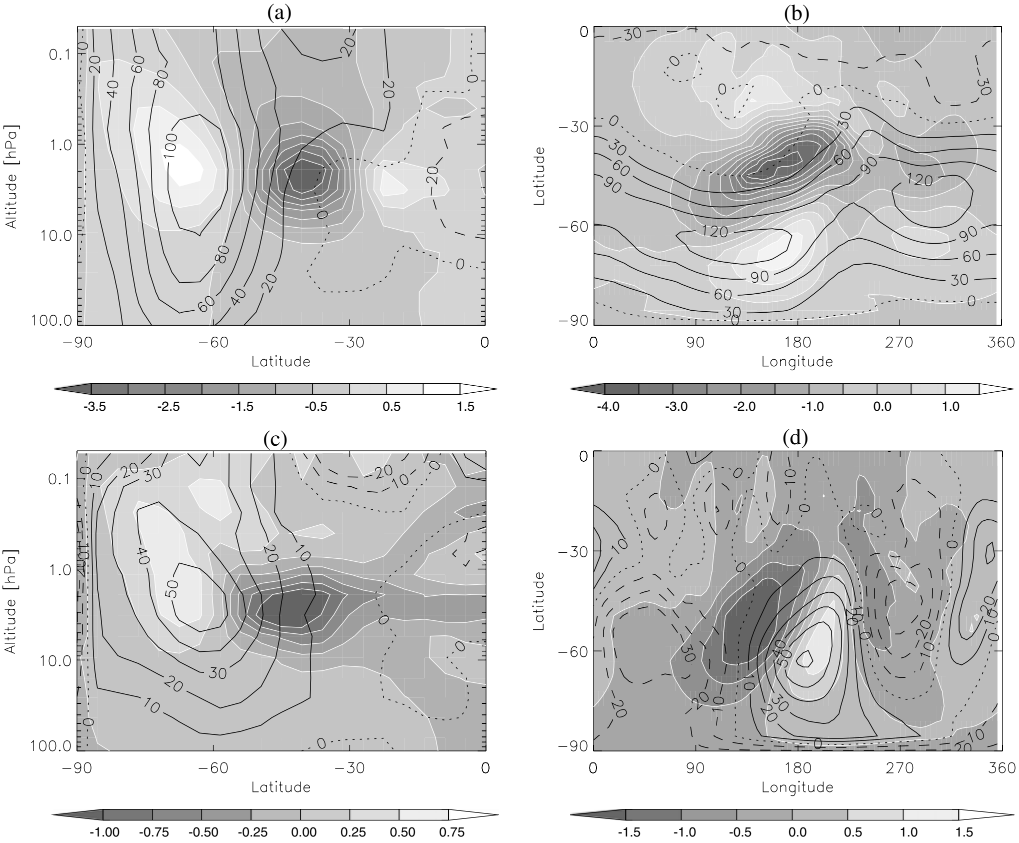

Figure 7. 'Observed' (a, b) zonal and (c, d) meridional winds (m s$^{-1}$) evolved from realistic initial conditions (black contours), and the difference between the true evolution and the control run (white contours with shading). (a) and (c) are vertical sections at 180° longitude, and (b) and (d) are horizontal sections at 1.9 hPa.

Despite these notable differences in the response, the technique can capture the GWD field as shown in Fig. 8 using the cost function defined in (14) with 25 minimization iterations. This effective estimation is due to the adjoint model; it can trace back along the trajectory and identify the right place where the source or sink of momentum is occurring, even when the effects of this forcing are being advected and deformed by the mean flow. As in the idealized cases, the technique can estimate the true drag. However it takes more iterations to achieve the same accuracy; in Fig. 8 errors are larger than Fig. 2 using the same number of minimization iterations. This happens because the realistic flow conditions lead to a more complex relationship between the drag field and the response to the drag, particularly involving advection and shear by the control flow. Consequently the system is less well conditioned than in the idealized case.

Diagnostics like that shown in Fig. 4 (see Fig. 9) show that, even in realistic flow conditions, the sensitivity of the model state to the drag remains linear to an excellent approximation. Thus, nonlinearity is not a limitation of the technique.

### (*b*) *Radiation*

The adjoint model only represents the adiabatic processes. There are two reasons which lead us to approximate the gradient of the cost function in this way.

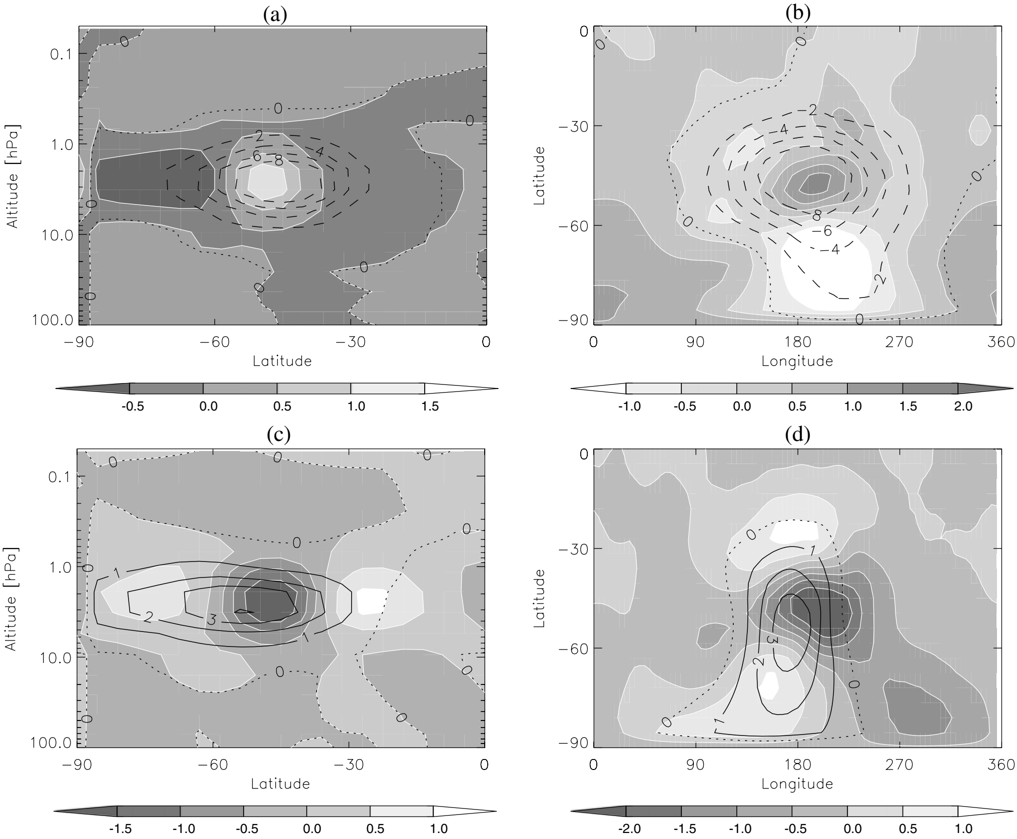

Figure 8. Estimated (a, b) zonal and (c, d) meridional gravity wave drag in realistic flow conditions (black contours) and errors in the estimation (white contours with shading), both m $\mathrm{s}^{-1}\mathrm{day}^{-1}$. (a) and (c) are vertical sections at 180° longitude, and (b) and (d) are horizontal sections at 1.9 hPa.

Firstly, the length of the assimilation window is much smaller than the radiative time-scale. Hence, the cost function must have only small sensitivity to radiative processes on this time-scale. The second reason is a practical one; the large number of operations that would be needed to calculate the adjoint of radiative processes would require much more computational resources, which could restrict the application of the technique.

The proposed alternative to avoid this limitation is to use a hybrid method, where the radiation processes are only taken into account in the forward model within the assimilation module to calculate the trajectory and initial conditions of the backward integration. In this case, if the dynamical gradient of the cost function is a good approximation to the full one, the ASDE should converge towards the true solution. Note this approximation does not change the solution, i.e. the optimal wave drag that minimizes $J$; it may only influence the convergence process. Such hybrid methods, in which the cost function gradient is approximated without using the full exact adjoint of the nonlinear forward model, are often used for data assimilation (Lawless *et al.* 2003 and references therein).

The experiment of section 5(a) was repeated, this time including $\dot{\theta}$ calculated using the radiative scheme, both in the 'truth' integration and in the ASDE forward

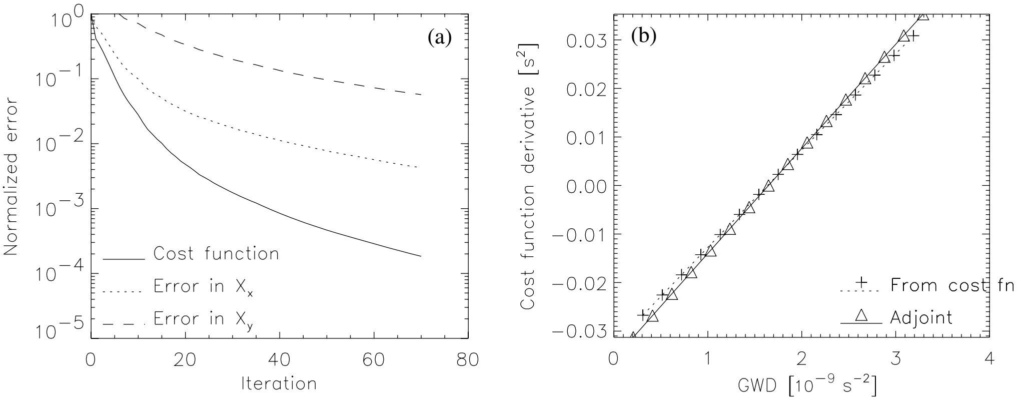

Figure 9. Performance of the technique with radiative processes: (a) is as Fig. 4(a), and (b) is as Fig. 3(b) in one particular search direction. In (b), note the slight differences between the exact derivative and the approximated one; these differences do not produce a change in the root.

integrations, but neglecting the sensitivity of $\dot{\theta}$ in the adjoint calculations. This hybrid method presented very good convergence rates, as shown in Fig. 9(a). For all the iterations, the cost function diminishes and so does the error in the drag estimation. The reason for this success is that the sensitivity of $\sigma$ to $\dot{\theta}$ and sensitivity of $\dot{\theta}$ to $\sigma$ are both small on the assimilation timescale. Therefore sensitivity of $\sigma$ to $\sigma$, via radiation, is tiny. Furthermore, as seen in Fig. 9(b), the effects of radiative processes make no noticeable change in the root of the derivative, which is the same as the one calculated with the hybrid method. This feature is found in all the search directions.

## 6. Estimation with a non-perfect radiative scheme

So far we have been working with a perfect model, and therefore all the differences between the observations and the control state were produced by the GWD. However, in reality there may be differences that are not produced by the drag but by other physical processes that are not perfectly represented in the numerical model, notably radiative processes. In this case ASDE will interpret the mismatch as coming from a GWD and will try to estimate a drag that minimizes those differences, leading to errors in the estimated drag.

To investigate the errors in the estimated drag that could result from errors in the model radiation scheme, we carried out a twin experiment in which the truth is calculated taking into account radiative processes via the standard radiative scheme, while in the ASDE system all radiative heating rates were multiplied by 1.2 to simulate a 20% bias. Since the initial state is far from equilibrium, the radiative heating rates, and hence the imposed errors, are actually very large.

The errors in the estimated drag are dominated by the radiation errors. Figure 10 shows the errors in the estimated GWD. The errors are largest in the mesosphere, where the radiative heating bias leads to biggest errors in $\sigma$ and $Q$ (not shown). These values scale almost linearly with the radiation errors, so they give an idea of the magnitude of drag errors that can be expected for a given radiation error.

Other tests have shown what happens if the ASDE system has a global-scale radiation error of one sign, localized in the vertical. It leads to a global-scale bias in $\sigma$ and hence $Q$ in the control integration. When the cost function involves $Q$, the system

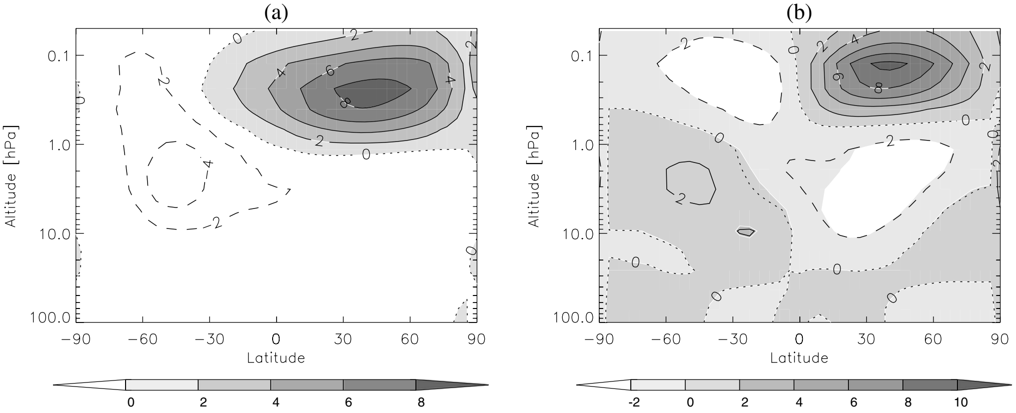

Figure 10. Zonally averaged fields of (a) zonal and (b) meridional gravity wave drag for the overestimated radiation simulation. Black contours show estimated GWD, and white contours with shading show the errors in the estimation (both m $s^{-1}day^{-1}$).

tries to reduce $Q$ errors by moving air polewards or equatorwards. The estimated drag thus contains a global-scale poleward or equatorward meridional component localized in the vertical (similar patterns are also visible in Fig. 10). This error is entirely found in the divergent component of the drag, not in the rotational component. The characteristic, and unrealistic, pattern of drag makes such radiation errors quite easy to identify in practice, and indeed has helped to identify and correct errors in the bottom-boundary upward radiative fluxes used in the model.

## 7. Conclusions

Twin experiments have shown that a technique based on variational data assimilation can be used to estimate gravity wave drag from a time series of global middle atmosphere data. The technique does not rely on zonal averages or long time averages, and so is able to estimate the three-dimensional distribution of both zonal and meridional drag components and their day-to-day variations.

The technique is computationally affordable. First, for an assimilation window of one day and for realistic drag amplitudes, the dependence of the model state on the drag is linear to an excellent approximation. Second, for a suitable choice of control variables and cost function, the minimization problem is well conditioned. These two factors mean that, for practical purposes, the iterative minimization converges in about 10 iterations for resting initial conditions, or about 20 iterations for strong realistic winds, with errors in the estimated drag of order 1 m $s^{-1}day^{-1}$. This degree of accuracy would certainly give useful information in the mesosphere and upper stratosphere, where drag amplitudes are large; in the lower stratosphere, particularly in the quasi-biennial oscillation region where drag amplitudes are much smaller, more iterations might be needed, along with an average over several assimilation windows to improve ‘signal-to-noise’, even if the observed data are sufficiently accurate.

To account for radiative processes we use a hybrid method, which considers radiative processes in the forward model evolution but neglects their effects in the adjoint model evolution. This approximation to the adjoint model gives an approximate directional gradient of the cost function that is very close to the exact gradient, and a zero of the directional gradient that is extremely close to the exact zero.

The technique can give useful information even with observations of only one variable. In particular, a series of experiments was performed where only pseudo-density, which contains only temperature information, was the observed variable. It was shown that there is a reasonable convergence of the rotational GWD component for midlatitudes. However, in the tropics the pseudo-density conservation equation is almost uncoupled from the momentum equations, so pseudo-density contains almost no information on the momentum forcing.

For a given number of iterations, the rotational component of the drag is found to be better estimated than the divergent component, even with perfect observations. This result can be understood in terms of the relationship between the drag and the flow response, affecting the conditioning of the two parts of the estimation problem, and in terms of the information lost as large-scale gravity waves forced by the drag propagate away and are dissipated. Moreover, errors in model radiation tend to manifest themselves as errors in the divergent component of the drag. For practical purposes it is most important that the rotational component of the drag be well estimated, since even a transient rotational drag will produce a long-term response in the balanced, rotational part of the flow, whereas a transient divergent drag will produce only a transient large-scale gravity wave response.

This paper has focused on the description and technical aspects of the new technique. A second part will apply the technique to estimate gravity wave drag from Met Office analyses, and discuss the additional sources of error that arise when using the technique with real-world data.

## APPENDIX

### *Linear response to the gravity wave drag*

The knowledge of the response to the forcing may give valuable information to obtain an optimal estimation of the real GWD. The determination of a suitable **R** matrix to get a well-conditioned minimization problem, and the choice of best variables to observe, depend on the characteristics of the solution.

The adjustment of the atmosphere to external forcing has been extensively studied in the literature (e.g. Blumen 1972; Zhu and Holton 1987; Weglarz and Lin 1998). As has already been noted, the analysis of the problem is clearer using the vorticity and divergence equations, therefore the equations of the dynamical model we use, (6)–(8), are especially suitable for this study. We assume an isothermal background state on an $f$-plane. Linearizing about a state of rest gives

$$\{(\partial_{tt}^2 + f^2)\mathcal{H} + \nabla^2\}\partial_t(\overline{\sigma}^{-1}\sigma') = -\mathcal{H}(\partial_t X_\delta + f X_\zeta) \qquad \text{(A.1)}$$

$$\{(\partial_{tt}^2 + f^2)\mathcal{H} + \nabla^2\}\partial_t\zeta' = \nabla^2 X_\zeta + \mathcal{H}(\partial_{tt}^2 X_\zeta - f\partial_t X_\delta) \qquad \text{(A.2)}$$

$$\{(\partial_{tt}^2 + f^2)\mathcal{H} + \nabla^2\}\delta' = \mathcal{H}(\partial_t X_\delta + f X_\zeta), \qquad \text{(A.3)}$$

where $\mathcal{H} = (g\overline{\sigma})^{-1}\partial_\theta(\overline{\rho}\theta\partial_\theta)$.

The solution can be expressed in Fourier components with the fields given by

$$(\sigma'/\overline{\sigma}, \zeta', \delta') = \{\widehat{\sigma}(t)/\overline{\sigma}, \widehat{\zeta}(t), \widehat{\delta}(t)\}\exp\{\mathrm{i}(kx + ly) + (1/2H + \mathrm{i}m)z\}. \qquad \text{(A.4)}$$

The solutions of the homogeneous equations are free inertia–gravity waves which satisfy the dispersion relationship

$$\omega^2 = f^2 + \frac{(k^2 + l^2)N^2}{(1/4H^2) + m^2} \qquad \text{(A.5)}$$

and a geostrophic mode of frequency, $\omega = 0$.

Expressing the forcing in Fourier components

$$(X'_\delta, X'_\zeta) = (\widehat{X}_\delta, \widehat{X}_\zeta) \exp\{\mathrm{i}(kx + ly) + (1/2H + \mathrm{i}m)z\}, \tag{A.6}$$

the forced solution is

$$\frac{\widehat{\sigma}}{\overline{\sigma}} = -\frac{f}{\omega^2}\widehat{X}_\zeta t - \frac{\widehat{X}_\delta}{\omega^2}\{1 - \cos(\omega t)\} + \frac{f}{\omega^3}\widehat{X}_\zeta \sin(\omega t), \tag{A.7}$$

$$\widehat{\delta} = \frac{f\widehat{X}_\zeta}{\omega^2}\{1 - \cos(\omega t)\} + \frac{\widehat{X}_\delta}{\omega} \sin(\omega t), \tag{A.8}$$

$$\widehat{\zeta} = \left(1 - \frac{f^2}{\omega^2}\right)\widehat{X}_\zeta t - \frac{f\widehat{X}_\delta}{\omega^2}\{1 - \cos(\omega t)\} + \frac{f^2}{\omega^3}\widehat{X}_\zeta \sin(\omega t). \tag{A.9}$$

The potential vorticity perturbation is given by $Q' = (\zeta' - f\sigma')/\overline{\sigma}$, and using (A.7) and (A.9) we find

$$Q' = X_\zeta t/\overline{\sigma}. \tag{A.10}$$

The forcing in the vorticity equation produces a geostrophically balanced growing anomaly in $Q$, $\zeta$ and $\sigma$, along with some inertia–gravity waves. The forcing in the divergence equation produces steady $\zeta$ and $\sigma$ anomalies along with some inertia–gravity waves. Note, in particular, that $Q$ is affected only by $X_\zeta$, not $X_\delta$.

On very short time-scales, $\omega t \ll 1$, we find

$$\sigma' = 0, \tag{A.11}$$

$$\zeta' = X_\zeta t, \tag{A.12}$$

$$\delta' = X_\delta t. \tag{A.13}$$

In this case the response to the forcing is particularly simple; it is local to the forcing and independent of the spatial scale of the forcing. This allows the matrix $\widehat{\mathbf{M}}$ to be written down explicitly.

## REFERENCES

Alexander, M. J. and Rosenlof, K. H. 2003 Gravity-wave forcing in the stratosphere: Observational constraints from the upper-atmosphere research satellite and implications for parameterization in global models. *J. Geophys. Res.*, **108**(D19)**,** doi: 10.1029/ 2003JD003373

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., Holton, J. R., Alexander, M. J., Hirota, I., Horinouchi, T., Jones, D. B. A., Kinnersley, J. S., Marquardt, C., Sato, K. and Takahashi, M. 2001 The quasi-biennal oscillation. *Rev. Geophys.*, **39,** 179–229

Blumen, W. 1972 Geostrophic adjustment. *Rev. Geophys.*, **10,** 485–528

Errico, R. M. 1997 What is an adjoint model? *Bull. Am. Meteorol. Soc.*, **78,** 2577–2591

Giering, R. and Kaminski, T. 1998 Recipes for adjoint code construction. *ACM Trans. Math. Software*, **24,** 437–474

Gregory, A. R. 1999 'Numerical simulations of winter stratospheric dynamics'. PhD thesis, University of Reading, UK

Hamilton, K. 1983 Diagnostic study of the momentum balance in the northern hemisphere winter stratosphere. *Mon. Weather Rev.*, **111,** 1434–1441

Hines, C. O. 1997 Doppler spread parametrization of gravity-wave momentum deposition in the middle atmosphere. Part 1: Basic formulation. *J. Atmos. Solar–Terr. Phys.*, **59,** 371–386

Klinker, E. and Sardeshmukh, P. 1992 The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci.*, **49,** 608–627

Lawless, A. S., Nichols, N. K. and Ballard, S. P. 2003 A comparison of two methods for developing the linearization of a shallow-water model. *Q. J. R. Meteorol. Soc.*, **129,** 1237–1254

Lindzen, R. S. and Holton, J. 1968 A theory of the quasi-biennial oscillation. *J. Atmos. Sci.*, **25,** 1095–1107

Lindzen, R. S. 1981 Turbulence and stress owing to gravity wave and tidal breakdown. *J. Geophys. Res.*, **86,** 9707–9714

Marks, C. J. 1989 Some features of the climatology of the middle atmosphere revealed by Nimbus 5 and 6. *J. Atmos. Sci.*, **46,** 2485–2508

Navon, I. M. and Legler, D. M. 1987 Conjugate-gradient methods for large-scale minimization in meteorology. *Mon. Weather Rev.*, **115,** 1479–1502

Palmer, T. N., Shutts, G. J. and Swinbank, R. 1986 Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through on orographic gravity wave drag parametrization. *Q. J. R. Meteorol. Soc.*, **112,** 1001–1039

Scaife, A. A., Butchart, N., Warner, C. D. and Swinbank, R. 2002 Impact of a spectral gravity wave parameterization on the stratosphere in the Met Office unified model. *J. Atmos. Sci.*, **59,** 1473–1489

Shine, K. 1987 The middle atmosphere in the absence of dynamical heat fluxes. *Q. J. R. Meteorol. Soc.*, **113,** 603–633

1989 Sources and sinks of zonal momentum in the middle atmosphere diagnosed using the diabatic circulation. *Q. J. R. Meteorol. Soc.*, **115,** 265–292

Shine, K. P. and Rickaby, J. A. 1989 Solar radiative heating due to absorption by ozone. Pp. 597–600 in *Ozone in the atmosphere*. Eds. R. D. Bojkov and P. Fabian, Deepack Publishing, Hampton, USA

Swinbank R. and O'Neill, A. 1994 A stratosphere–troposphere data assimilation system. *Mon. Weather Rev.*, **122,** 686–702

Thuburn, J. 1996 Multidimensional flux-limited advection schemes. *J. Comput. Phys.*, **123,** 74–83

Thuburn, J. and Haine, T. 2001 Adjoints of non-oscillatory advection schemes. *J. Comput. Phys.*, **171,** 616–631

Vukićević, T., Steyskal, M. and Hecht, M. 2001 Properties of advection algorithms in the context of variational data assimilation. *Mon. Weather Rev.*, **129,** 1221–1231

Warner, C. D. and McIntyre, M. E. 1996 On the propagation and dissipation of gravity wave spectra through a realistic middle atmosphere. *J. Atmos. Sci.*, **53,** 3213–3235

Weglarz, R. P. and Lin, Y. L. 1998 Nonlinear adjustment of a rotating homogeneous atmosphere to zonal momentum forcing. *Tellus*, **50,** 616–636

Zhu, X. and Holton, J. R. 1987 Mean fields induced by local gravity-wave forcing in the middle atmosphere. *J. Atmos. Sci.*, **44,** 620–630