An Optimal-Control/Adjoint-Equations Approach to Studying the Oceanic General Circulation

ELI TZIPERMAN

Isotope Department, The Weizmann Institute of Science, Rehovot, Israel

WILLIAM CARLISLE THACKER

Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida

(Manuscript received 22 August 1988, in final form 14 April 1989)

ABSTRACT

An efficient procedure is presented for analyzing oceanographic observations with the aid of a general circulation model. Poorly known model parameters, such as eddy-mixing coefficients, surface forcing and tracer boundary fluxes, can be calculated by fitting model results to observations. Optimal estimates for all model fields, including the observed ones, can then be computed by running the model with the best-fit values of the calculated parameters. Information about the resolution and the error-covariances of the model parameters can be computed. This information is shown to be very valuable for critically evaluating how well the data determine the parameter's values. An adjoint model, similar in structure to the numerical model, uses information on model–data misfit to improve estimates of the unknown model parameters, and improve the fit to observations. The procedure is illustrated using simulated data and a simple, barotropic, nonlinear, quasi-geostrophic model. Examples are discussed in which friction parameters, wind forcing, and the steady-state circulation are determined from simulated vorticity and streamfunction observations.

1. Introduction

Any analysis of oceanic data uses a model in the hope of explaining the data and improving the less well-known aspects of the model. A major challenge confronting numerical modelers and observational physical oceanographers is to analyze the large quantities of oceanographic data of many types with the aid of complex numerical general circulation models (GCMs). Numerical models have become quite sophisticated in recent years, but they must still be given many poorly known inputs such as eddy-mixing coefficients, surface forcing by heat and momentum fluxes, and tracer boundary fluxes. There is normally no direct information on many of these input parameters in oceanic measurements, and numerical models are presently not capable of using the existing data for the interior temperature, salinity, and/or currents to deduce the unknown model inputs. The oceanic data itself is noisy and incomplete in spatial and temporal coverages, and one would like to be able to use the numerical model to obtain a better estimate of the observed fields while improving the model itself.

The determination of a model's input parameters is usually referred to as an inverse problem, the direct problem being the simulation of the ocean assuming values for the inputs to be given. Inverse models, unlike numerical GCMs, are capable of using various types of data in order to calculate mixing coefficients and velocity field, but presently suffer from some limitations as well. The model is often formulated as a set of linear equations relating data and unknown parameters, written in matrix form and solved by methods such as singular value decomposition, or linear programming. Computations are generally carried out in such a way that the matrices, having dimensions of (number of unknowns × number of equations), must be stored in the computer memory. One finds that the computational requirements with this approach rapidly increase with model complexity, and in particular the storage of the matrices becomes a limiting factor. As a result, the models have been limited to low spatial resolution (for box models, Wunsch 1978) or to simple local dynamics disregarding boundary conditions and mass continuity (Olbers et al. 1985).

A specific example may clarify the problem of large matrices. For a GCM of the North Atlantic, with a 30 by 30 horizontal grid, eight vertical levels, and five fields (two components of horizontal velocity, temperature, salinity, and some passive tracer), the number of unknowns, not counting momentum and heat fluxes at the surface, is $30 \times 30 \times 8 \times 5 = 36\,000$. This is also the number of model equations for these unknowns, so that the dimension of the matrix involved in using

Corresponding author address: Dr. Eli Tziperman, The Weizmann Institute of Science, Isotope Department, 76100 Rehovot St., Israel.

the aforementioned methods would be 36 000 $\times$ 36 000, or more than 10^9 numbers. This is obviously unacceptable for a model of such modest resolution, and the storage requirements for a full oceanic GCM would be formidable.

An inverse problem can also be formulated as an optimization problem, in which case nonlinear models can also be accommodated (Schröter and Wunsch 1986). A cost function measuring the model-to-data distance is minimized with the model equations treated as constraints. The minimum is computed by solving a large system of coupled nonlinear equations for all model variables (velocities, temperature, wind, etc.), which is not a trivial task. With no information on the gradient of the cost function with respect to all model variables and inputs, the search for the minimum of the cost function with respect to its many arguments is very inefficient and can be very slow. In addition, the optimization algorithm must be based on the Hessian matrix of second partial derivatives of the cost function (e.g., in Newton's method), so it is again necessary to confront a matrix whose dimensions are the square of the number of unknown model parameters.

In order to work with large GCMs, one needs an efficient and storage effective way of computing the gradient of the cost function for an oceanic inverse problem involving a GCM. Then attention should be focused on finding gradient-based algorithms that can find the minimum efficiently. This is the approach taken here.

An efficient way to calculate the gradient is provided by the adjoint-equations technique, which is widely used in engineering (Hasdorff 1976). It has been proposed for use in meteorological data assimilation (Le Dimet and Talagrand 1986) and also for oceanographic data assimilation (Thacker and Long 1988). The basic idea is that the gradient can be computed using an adjoint model, which integrates the adjoints of the partial-differential equations that define the numerical model. The first and most significant advantage of this approach is that it allows the gradient vector to be accurately evaluated for a computational expense roughly equivalent to a single model run. This advantage is easily appreciated by comparing this approach with that of evaluating the gradient by finite differences, which would require additional model runs for each of the many variables. In addition, the adjoint technique provides a way to avoid having to work with large matrices, and the resulting storage problems.

The purpose of this paper is to demonstrate how the adjoint technique makes feasible the use of a full size GCM for the analysis of oceanographic data. The adjoint technique serves as an efficient and powerful inverse method: unknown model parameters are calculated by fitting the model results to the data, and an optimal estimate of the observed fields is obtained as well. A solution for the model parameters is incomplete, however, without error and resolution information, and these can also be calculated using the present approach. The method allows the use of steady data such as averaged temperature, salinity and current-meter measurements, as well as time-dependent data such as transient tracers or observations of seasonal processes. A simple barotropic nonlinear quasi-geostrophic model and simulated data are used here to demonstrate the method. The steady circulation, wind-forcing and friction parameters are calculated from vorticity or streamfunction observations, and the importance of the resolution information is demonstrated.

The present approach is closely related to the model fitting discussed by Thacker and Long (1988), who used a simple equatorial wave model to calculate optimal initial conditions in order to fit simulated observations. Thacker (1987) also discussed the calculation of the covariance matrix for the calculated parameters, and this is further developed here, extended to include resolution information, and demonstrated by a specific example. Wunsch (1988) has discussed the application of optimal control methods to the oceanic transient-tracers problem, demonstrating their advantages in the case of time-dependent data. He used a matrix formulation of the optimal control technique, in principle equivalent to the one used here, although difficult to apply to larger GCMs, as discussed above. Variational data assimilation (Le Dimet and Talagrand 1986), adjoint sensitivity analysis (Cacuci 1981; Hall and Cacuci 1983), and other variational methods (Bennett and McIntosh 1982), are all based on similar approaches to that discussed here. In addition, although using a different methodology, Schröter and Wunsch (1986) have used the same QG model we use here, and have asked similar questions. The major difference between their work and the present one is in the methodology used. They have emphasized the discussion of the results, ignoring questions of efficiency and feasibility of using the method with larger realistic GCMs, while as shown above, one has to take care of these questions in order to be able to use even a very simple oceanic GCM with oceanographic data. Some of the specific results concerning the QG model, as well as the discussion of the possibilities of using a numerical model within an inverse procedure, are to some extent similar in the two works.

Given a set of measurements of various quantities related to the general circulation, the adjoint approach proceeds as follows: A cost function is defined, measuring the distance between model results and observations. The cost function is therefore a function of both the observations and the unknown model parameters. Given an initial guess for the model parameters, the numerical model is used to calculate the value of the cost function. An adjoint numerical model is then used to calculate the gradient of the cost function with respect to the many unknown model parameters. Next, an optimization algorithm (e.g., conjugate gradient) uses the gradient information to obtain a new guess

for the parameters, reducing the value of the cost function. Several such iterations are needed to obtain the minimum value of the cost function, where model results and observations are as close as allowed by the level of measurements noise. The optimal estimate for the parameters is that corresponding to the minimum value of the cost function. Having obtained the estimate for the unknown parameters, the Hessian matrix can be used to obtain error and resolution information that is necessary in order to evaluate the computed solution.

Section 2 describes the methodology, including the quasi-geostrophic model description (section 2*a*), the definition of the cost function (section 2*b*), the derivation of the adjoint equations and the optimization procedure (section 2*c*), and the calculation of error and resolution information from the Hessian matrix (section 2*d*). Some specific results of the quasi-geostrophic model runs are given in section 3, and we conclude in section 4.

2. Method

a. Model, data and problem definition

The model used here to demonstrate this optimal-control/adjoint-equations approach to studying the general circulation is based on the nonlinear, quasi-geostrophic, barotropic vorticity equation, in a closed rectangular domain. The nondimensional equation is (Veronis 1966; Pedlosky 1979):

$$\nabla^2\psi_t + \psi_x + RJ(\psi, \nabla^2\psi) = -\epsilon_b\nabla^2\psi + \epsilon_h\nabla^4\psi + \text{curl}\tau(x, y), \quad (1)$$

where (x, y) and (u, v) are the (east, north) coordinates and velocity components respectively, $0 \leqslant x, y \leqslant 1$, ψ is the streamfunction, $\psi_x = v$, $-\psi_y = u$, and the nondimensional friction parameters and Rossby number are given by

$$\epsilon_h = \frac{K_h}{\beta L^3}, \quad \epsilon_b = \frac{K_b}{\beta DL}, \quad R = \frac{W}{\beta^2 L^3 D},$$

where $\beta = df/dy$ is the gradient of the Coriolis parameter, D is the depth of the rectangular basin, L is the horizontal scale of the basin, W is the magnitude of the wind-stress, and K_b and K_h are coefficients of bottom and horizontal friction. Two boundary conditions are needed at each point on the boundary, due to the horizontal friction term (Pedlosky 1979). The no-flux and no-stress boundary conditions are used:

$$\psi = 0; \quad \zeta \equiv \nabla^2\psi = 0, \quad (2)$$

with ζ denoting vorticity. The finite-difference model is described in appendix A; starting with an initial vorticity field, the streamfunction is found by solving a Poisson equation, and then the vorticity is advanced in time. Computations are on a grid of I uniformly spaced grid points in the x-direction and J in the y-direction.

The steady state streamfunction and the forcing curlτ, at all grid points, as well as the friction parameters ϵ_b and ϵ_h, are regarded as unknowns, which are to be calculated by fitting the model to observations. In the examples described below, simulated observations were used. These were obtained by running the model to steady state, using $\text{curl}\tau = -\sin(\pi x)\sin(\pi y)$, $\epsilon_b = 0.05$, $\epsilon_h = 0.0001$ and $R = 0.01$, and saving the final streamfunction or vorticity to be used as data. Similar values were used by Schröter and Wunsch (1986), and by Veronis (1966). At these values the model is fairly nonlinear, and the corresponding steady-state solution for ψ and ζ is shown in Fig. 1.

b. The cost function

Given the model and the observations defined in the previous subsection, a cost function measuring the fit of the model results to the observations now needs to be chosen. We start by defining the cost function for the quasi-geostrophic example used in this work, and then make a few more general comments.

Suppose that the model (A1) is given some initial conditions for the vorticity ζ_{ij}^0, and is stepped forward to calculate ψ_{ij}^0, ζ_{ij}^1, and ψ_{ij}^1, where ζ_{ij}^n and ψ_{ij}^n denote the vorticity and streamfunction at horizontal location (i, j), and time level n. If ζ_{ij}^0 is in fact the steady-state solution, the difference between the initial conditions and the solution after one time step should vanish. In addition, the model-to-data fit requires the difference between the model solution and the observations to be as small as possible. A cost function is therefore defined, measuring the departure of the model solution both from steady state and from the observations. For this example assume that observations $\hat{\psi}$ and $\hat{\zeta}$ of streamfunction and vorticity are available at every grid point, realizing that such datasets are highly artificial. The cost function:

$$\mathbf{J}(\text{curl}\tau_{ij}, \epsilon_b, \epsilon_h, \zeta_{ij}^0) = \sum_{i,j} [C_{ij}^{(1)}(\psi_{ij}^0 - \hat{\psi}_{ij})^2 + D_{ij}^{(1)}(\psi_{ij}^1 - \psi_{ij}^0)^2 + C_{ij}^{(2)}(\zeta_{ij}^0 - \hat{\zeta}_{ij})^2 + D_{ij}^{(2)}(\zeta_{ij}^1 - \zeta_{ij}^0)^2]. \quad (3)$$

is a function of the unknown wind-stress curl and friction parameters via ζ_{ij}^1, which must satisfy the model equations. Observation errors at different locations are assumed to be uncorrelated, so $C_{ij}^{(k)}$ can be taken to be elements of the inverse error-covariance matrix of the observations (Thacker 1987). Similarly, the coefficients $D_{ij}^{(k)}$ $(k = 1, 2)$ can be interpreted as elements of the inverse error-covariance matrix of bogus observations that indicate no time change in the streamfunction and vorticity.

Because the simulated observations are completely consistent with an exact steady-state model solution,

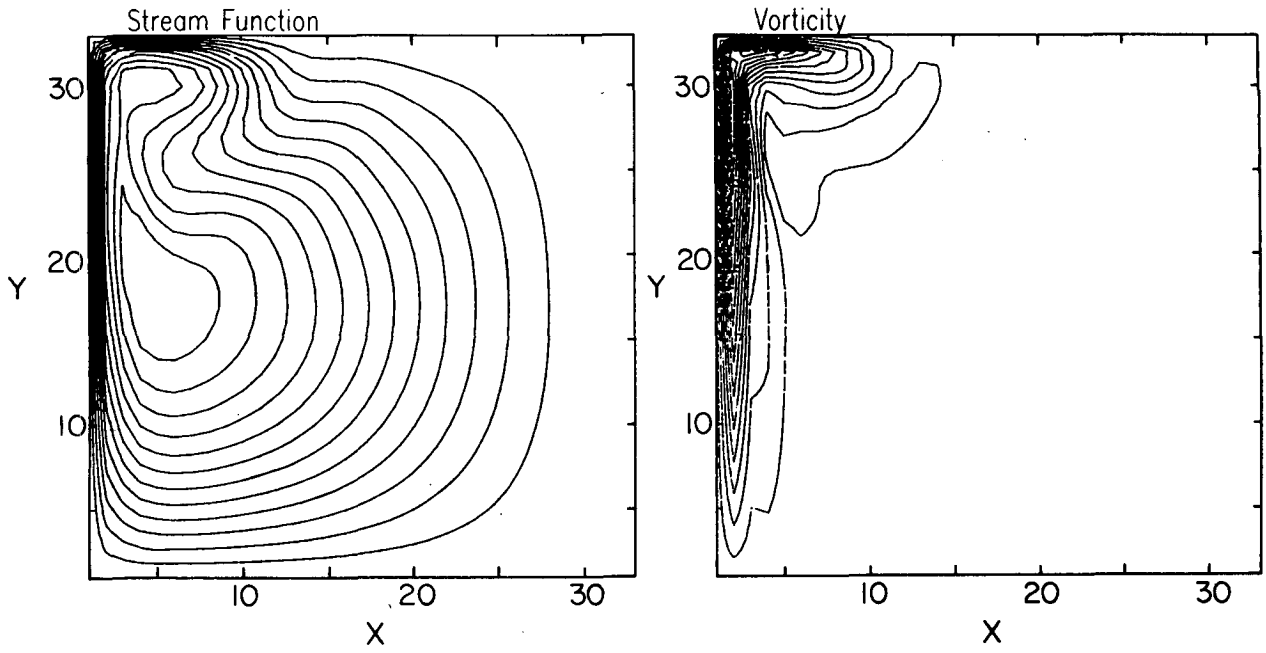

FIG. 1. The steady-state solution for streamfunction ψ and vorticity ζ, used as simulated observations. The parameters used to obtain this solution were $R = 0.01$, $\epsilon_b = 0.05$, $\epsilon_h = 0.0001$, $\text{curl}\tau = -\sin(\pi x)\sin(\pi y)$.

the optimal value of the cost function is $\mathbf{J} = 0$. The unknown parameters are then expected to be at their "true" values, and the initial conditions ζ_{ij}^0 found by the optimization are the exact steady-state solution. In a more realistic situation the data will not be consistent with a steady-state model solution; the minimum of $\mathbf{J}$ can be expected to be positive, and the optimal initial conditions can be expected to differ both from the observations and from a perfectly steady state.

With noisy observations, one might need to add a penalty term [for example, the square of the Laplacian of the vorticity $(\nabla^2\zeta)^2$ to the cost function, to ensure a spatially smooth solution (Bennett and McIntosh 1982; Thacker 1988)]. The terms proportional to $D_{ij}^{(k)}$ can be thought of as penalty terms requiring the solution to be in a steady state. In the limit that these coefficients get arbitrarily large, the initial conditions will correspond exactly to a steady solution of the time-dependent model. The ratio of $C_{ij}^{(k)}$ to $D_{ij}^{(k)}$ determines how near to a steady state the solution will be. In order to examine the performance of the adjoint procedure under the simplest possible circumstances, no noise was added to the simulated observations in the runs described in section 3. Nevertheless, as will be seen below, some interesting results and difficulties are found.

Let us make a few general comments concerning the general application of this approach to the general oceanic circulation. In the place of the barotropic vorticity equation, which is used here as an example, there might be a primitive-equation model with capability of simulating the transport of passive tracers, and in the place of the artificial, model-generated streamfunction and vorticity data, there would be actual oceanographic observations. Consider first the case of time-dependent data such as transient-tracers data, or, when modeling the seasonal variability of the ocean, data that resolves the seasonal variability. For transient tracers, every run of the forward model must span the time over which observations are available, and the cost function will contain appropriate terms measuring the distance between tracer observations at different times, and model results for the same tracers at the times of the observations. The velocities, temperature and salinity will still be required to be in a steady state, using penalty terms similar to those multiplying the $D_{ij}^{(k)}$ in (3). When modeling the seasonal variability, one may require the solution to have a one-year cycle, using some appropriate terms in the cost function. Each iteration will then involve running the forward and adjoint models for one year.

Note also that terms maximizing and minimizing different quantities of interest may be added to the cost function in order to find the upper and lower bounds on these quantities consistent with the data and the model (Wunsch 1984; Schröter and Wunsch 1986).

With the definition (3) of $\mathbf{J}$, our inverse problem is to find the values of $\text{curl}\tau_{i,j}$, ϵ_b, ϵ_h and ζ_{ij}^0 that minimize the cost function $\mathbf{J}$, therefore giving the best fit of the model to the observations. The procedure used to minimize $\mathbf{J}$ and calculate the correct parameters is described in the following subsection.

c. Adjoint model, optimization procedure

The strategy for computing the minimum of the cost function is to use a gradient-based iterative algorithm, since the gradient can be computed at low cost using an adjoint model. To derive the adjoint model equations (Thacker and Long 1988), form a Lagrange function by adding to the cost function the finite-difference model equations (appendix A):

$$\mathbf{L} = \mathbf{J} + \sum_{i=2}^{I-1}\sum_{j=2}^{J-1}\sum_{n=0}^{N} \mu_{ij}^n[\nabla^2\psi_{ij}^n - \zeta_{ij}^n] + \sum_{i=2}^{I-1}\sum_{j=2}^{J-1}\sum_{n=1}^{N} \{\lambda_{ij}^n[(\zeta_{ij}^n - \zeta_{ij}^{n-1})/\Delta t + (\psi_{i+1,j}^{n-1} - \psi_{i-1,j}^{n-1})/\Delta x + RJ(\psi_{ij}^{n-1}, \zeta_{ij}^{n-1}) + \epsilon_b\zeta_{ij}^{n-1} - \epsilon_h\nabla^2\zeta_{ij}^{n-1} - \text{curl}\tau_{ij}]\}, \quad (4)$$

where N is determined by the number of time levels that appear in the data and therefore also in the cost function. For the cost function (3), data enter only at $n = 0$ and terms measuring departures from the steady state involve $n = 0$ and $n = 1$, so set $N = 1$ for the present example. In general, the boundary conditions should also appear as constraints with their own Lagrange multipliers; however, for prescribed values at the boundaries, it is simpler to exclude the boundary values from the set of unknowns. At the constrained minimum of $\mathbf{J}$, the Lagrange function $\mathbf{L}$ has a stationary point with respect to ψ_{ij}^n, ζ_{ij}^n, λ_{ij}^n, μ_{ij}^n, $\text{curl}\tau_{ij}$, ϵ_b, ϵ_h and the initial conditions ζ_{ij}^0. Stationarity with respect to the Lagrange multipliers, $\partial\mathbf{L}/\partial\lambda_{ij}^n = 0$ and $\partial\mathbf{L}/\partial\mu_{ij}^n = 0$, gives the original model equations, whereas stationarity to streamfunction and vorticity, $\partial\mathbf{L}/\partial\psi_{ij}^n = 0$ and $\partial\mathbf{L}/\partial\zeta_{ij}^n = 0$, gives a set of adjoint equations for the Lagrange multipliers. The finite difference form of the adjoint equations is given in appendix A.

An adjoint can also be derived from the continuous equation (1) (see appendix B), and is more conveniently presented this way:

$$\nabla^2\lambda_t + \lambda_x + R[J(\lambda, \nabla^2\psi) - \nabla^2 J(\lambda, \psi)] = \epsilon_b\nabla^2\lambda - \epsilon_h\nabla^4\lambda + \Phi, \quad (5)$$

where the forcing term Φ, which contains information about the model–data misfit, depends upon the specific definition of the cost function $\mathbf{J}$. Note that the frictional terms appearing in the adjoint equation are identical to those of the forward equation (1), except for an opposite sign; because the adjoint equation is integrated backward in time (Hall and Cacuci 1983; Thacker 1987), the frictional terms do not cause numerical instability, and the adjoint problem is well posed mathematically. The physical significance of the adjoint variable (λ) is in propagating information about the sensitivity of the cost function to variations in the physical variables (ψ) (Hall and Cacuci 1983). The adjoint variables collect information on the misfit of model and data through the adjoint forcing terms in (5), and are later used to calculate the gradient of the cost function with respect to the unknown parameters. A variety of discrete adjoint models could be derived from (5) using different discretization methods. However, the only one that would be the adjoint of the finite-difference model is the one obtained using the Lagrange multiplier technique; the others may introduce truncation errors into the gradient of the cost function.

The boundary conditions for the adjoint equations are derived directly from the Lagrange function (4). These turn out to be homogeneous:

$$\lambda_{ij}^n = \mu_{ij}^n = 0 \quad \text{on the boundaries.} \quad (6)$$

Note that there are no Lagrange multipliers associated with the boundary points in the definition of the Lagrange function (4); introducing them as boundary conditions is simply a convenience. Similarly, setting to zero all Lagrange multipliers outside the time interval over which the cost function $\mathbf{J}$ is defined results in homogeneous initial conditions:

$$\lambda_{ij}^{N+1} = \mu_{ij}^{N+1} = 0, \quad (7)$$

signifying that at the initial time for the adjoint run the adjoint variables do not contain any information on model–data misfit (Thacker and Long 1988; Hall and Cacuci 1983).

The next step is to use the Lagrange multipliers to evaluate the gradient of the cost function. Note the important fact that the partial derivatives of the Lagrange function with respect to the unknown model parameters, taken as if ψ_{ij}^n, ζ_{ij}^n, λ_{ij}^n, μ_{ij}^n, $\text{curl}\tau_{ij}$, ϵ_b, ϵ_h, and ζ_{ij}^0 are all independent, are equal to the corresponding derivatives of the original cost function, taken with only the model input parameters varying independently:

$$\frac{\partial\mathbf{L}(\psi_{ij}^n, \zeta_{ij}^n, \lambda_{ij}^n, \mu_{ij}^n, \text{curl}\tau_{ij}, \epsilon_b, \epsilon_h, \zeta_{ij}^0)}{\partial\alpha_l} = \frac{\partial\mathbf{J}(\text{curl}\tau_{ij}, \epsilon_b, \epsilon_h, \zeta_{ij}^0)}{\partial\alpha_l}, \quad (8)$$

where α_l, $l = 1, \cdots, L$ are the unknown inputs $\text{curl}\tau_{ij}$, ϵ_b, ϵ_h, and ζ_{ij}^0. Writing the components of the gradient of the cost function explicitly for all α_l:

$$\frac{\partial\mathbf{J}}{\partial\epsilon_b} = \sum_{i=2}^{I-1}\sum_{j=2}^{J-1}\sum_{n=1}^{N} \lambda_{ij}^n\zeta_{ij}^{n-1}$$

$$\frac{\partial\mathbf{J}}{\partial\epsilon_h} = \sum_{i=2}^{I-1}\sum_{j=2}^{J-1}\sum_{n=1}^{N} -\lambda_{ij}^n\nabla^2\zeta_{ij}^{n-1}$$

$$\frac{\partial \mathbf{J}}{\partial \operatorname{curl}\tau_{ij}} = \sum_{n=1}^{N} \lambda_{ij}^{n}$$

$$\frac{\partial \mathbf{J}}{\partial \zeta_{ij}^{0}} = -\lambda_{ij}^{0}/\Delta t, \tag{9}$$

where λ_{ij}^{0} does not multiply any equation in the cost function, and is obtained by stepping the adjoint equation an additional time step from $n = 1$ to $n = 0$ (see appendix A).

The above expressions allow the calculation of the gradient of the cost function with respect to the many parameters of the problem with only one forward run of the quasi-geostrophic model and one backward integration of the adjoint model. A calculation of the gradient by perturbation analysis (varying the parameters by a small amount and recalculating the cost function by forward runs) would require a number of forward model runs equal to the number of unknown parameters—a much less efficient procedure. This is also the basis for using the adjoint method for sensitivity analysis of large numerical models (Hall and Cacuci 1983).

A descent algorithm can be used to minimize the cost function according to the following procedure:

1) Start from an initial guess for α_l.
2) Run model forward in time, calculate the cost function.
3) Run the adjoint model backward in time.
4) Calculate $\nabla_\alpha \mathbf{J}$ using (9) with the solutions for ψ, ζ, λ and μ from steps 2 and 3.
5) Use an optimization procedure [conjugate gradient (Gill et al. 1981) was used here], with the information on the value and gradient of the cost function, to obtain a new set of values for the parameters, and return to step 2. Continue iterations until cost function is at its minimum value.

d. Conditioning, resolution and error analysis

In order for the solution for the unknown parameters to be complete, it needs to be supplemented with error estimates and resolution information, and these can be obtained through the eigenvalues and eigenvectors of the Hessian matrix. As explained in the Introduction, the Hessian for even simple GCMs may be too large to handle directly, and in fact this is one of the main motivations for using the adjoint method. Unfortunately, in order to analyze the solution of the inverse problem, one must again confront the Hessian. There are several ways one might approach the problem of dealing with the large Hessian in this context. Assuming that the Hessian matrix is not rapidly varying for elements representing spatially close variables, one could calculate the Hessian for only a subset of the variables (say wind and streamfunction at only every second grid point in our model), and analyze the resulting smaller matrix. Another possibility, which we have chosen to use in the examples given below, is to calculate the Hessian for a model of a coarser resolution. The resulting resolution and error information may not be strictly correct for the higher resolution model, but as will be seen below, it still gives the necessary information on the validity of the solution. By reducing the size of the Hessian for the purpose of resolution analysis, the amount of storage needed for the eigenvectors is also reduced. There are other methods of using sparsity of the full Hessian both to reduce storage and to make its calculation more efficient (Ypma 1987). We wish to emphasize, however, that dealing with the full or even reduced Hessian matrix is not necessary for obtaining the solution itself.

To calculate the Hessian matrix $\mathbf{G}$, we have used simple finite differencing of the first derivatives,

$$\frac{\partial^2 \mathbf{J}}{\partial\alpha_i \partial\alpha_j} \approx \left(\left.\frac{\partial \mathbf{J}}{\partial \alpha_i}\right|_{\alpha_j+\delta\alpha_j} - \left.\frac{\partial \mathbf{J}}{\partial \alpha_i}\right|_{\alpha_j} \right) \frac{1}{\delta\alpha_j}, \tag{10}$$

where the first derivatives $\partial \mathbf{J}/\partial\alpha_i$ are calculated from the forward and adjoint solutions. With L parameters, this requires L runs of the forward and adjoint models, equivalent to $2L$ forward time steps in the present example. Note that the computational cost required to calculate the Hessian matrix may be much larger for time dependent data, where more than a single time step is needed at each forward or adjoint model run (see section 2*b*). An alternative procedure for calculating the Hessian matrix for nearly linear models, although at a similar computational cost, is discussed by Thacker (1987).

Once the Hessian has been evaluated, its eigenvalues and eigenvectors are required for analyzing resolution and variance. This is a significant computational problem for such large matrices, so some attention should be given to it in the future. Here we simply used a NAG library routine (Numerical Algorithms Group 1984).

1) CONDITIONING

In order for the conjugate-gradient descent algorithm to work efficiently and to converge to the minimum of $\mathbf{J}$ in a few iterations, the Hessian matrix, $\mathbf{G} = \{\partial^2 \mathbf{J}/\partial\alpha_i\partial\alpha_j\}$, $i, j = 1, \cdots, L$ must be well conditioned. Because $\mathbf{G}$ is symmetric and positive (at least near the minimum), its eigenvalues are real and positive. If they are indexed from largest to smallest $\gamma_1 \geqslant \cdots \geqslant \gamma_L \geqslant 0$, the conditioning number is defined as the ratio γ_1/γ_L. If the conditioning number is large, numerical difficulties can be anticipated (Gill et al. 1981).

Consider the significance of the eigenvalues and eigenvectors of the Hessian matrix. Being a symmetric matrix, the Hessian may be written in terms of its eigenvalues and eigenvectors as $\mathbf{G} = \mathbf{V}\Gamma\mathbf{V}^{\mathrm{T}}$ where Γ is a diagonal matrix whose elements are the eigenvalues of

G ordered in decreasing magnitude, and the columns $\mathbf{V}_i$ of the matrix **V** are the corresponding eigenvectors of G. Suppose now that a small vector $\Delta\alpha = \epsilon\mathbf{V}_i$ is added to the solution vector α at the minimum of **J**, where $\partial\mathbf{J}/\partial\alpha_i' = 0$. The resulting change in the cost function is

$$\Delta\mathbf{J} \approx \Delta\alpha^{\mathrm{T}} G \Delta\alpha = \epsilon^2 \mathbf{V}_i^{\mathrm{T}}(\mathbf{V}\Gamma\mathbf{V}^{\mathrm{T}})V_i = \epsilon^2\gamma_i. \quad (11)$$

It can be seen that the largest change in the value of the cost function is in the direction of the eigenvectors corresponding to the largest eigenvalues. When searching for the minimum of the cost function, the conjugate-gradient descent algorithm prefers to search in these directions. As a result, these are the best determined directions in the parameter space, or in other words, the best determined linear combinations of the model parameters. The conjugate gradient algorithm will not search in directions in parameter space characterized by the very small eigenvalues, because there is no change of the cost function for these directions. As a result, there is no information about the linear combinations of parameters corresponding to the small eigenvalues.

Small or zero eigenvalues of the Hessian can result from lack of data, and one might want to add smoothing terms as bogus data in the cost function representing prior knowledge, in order to supplement the actual measurements (Thacker 1988). This is necessary because small eigenvalues mean the Hessian is ill-conditioned, and therefore convergence to the minimum of the cost function is very slow. After adding the bogus data, there may be enough information on all unknown parameters to improve the rate of convergence.

If the conditioning number is too big, it is possible to speed up convergence by transforming to a new set of unknowns $\alpha' = \mathbf{W}^{-1}\alpha$ for which the transformed Hessian $\mathbf{G}' = \{\partial^2\mathbf{J}/\partial\alpha_i'\partial\alpha_j'\} = \mathbf{W}^{\mathrm{T}}\mathbf{G}\mathbf{W}$ is better conditioned (Gill et al. 1981). The optimal parameters of the new unknowns can be computed in fewer iterations, and then the optimal values of the original unknowns are given by the inverse transformation. The strategy is to find an easily invertible transformation that will improve the conditioning. In many cases a simple scaling transformation (a diagonal **W** matrix) is sufficient, and proves very useful. Preconditioning may be necessary even when a nondimensionalized model equation such as (1) is used. As an example of a nondiagonal preconditioning transformation, note that the initial conditions sought by the optimization algorithm in the present example may be either the vorticity ζ_{ij}^0 or the stream function ψ_{ij}^0. It was found that vorticity initial conditions are preferable, as using streamfunction initial conditions resulted in a very noisy gradient of the cost function and slow convergence to the optimal solution. Thus the transformation from streamfunction to vorticity is a preconditioning transformation, which corresponds to the choice $\mathbf{W}^{-1} = \nabla^2$; on the other hand, if the initial conditions were already in terms of the vorticity, then transforming to streamfunction would make the problem more ill-conditioned. In general, however, the preconditioning transformation need not define variables with any simple physical interpretation. Inverse problems involving large GCMs are quite likely to be ill-conditioned, and efforts will be needed to find preconditioning transformations that will speed up convergence.

2) RESOLUTION

Once the optimal solution for the unknown parameters has been found, it is useful to construct a parameter resolution matrix, similar to that used when solving linear systems by singular value decomposition (Wiggins 1972; Wunsch 1978). the resolution matrix indicates which parameters are in fact resolved by the data. This may be necessary, for example, when using the model and its adjoint for the purpose of experiment design, where one may want to find out what parameters are resolved by a particular set of data, and how accurate measurements should be.

The idea behind the resolution matrix is that eigenvectors of the preconditioned Hessian with large eigenvalues are well-determined linear combinations of the model parameters, while those with small eigenvalues are poorly determined combinations (11). The data fail to resolve those linear combinations whose eigenvalues are too small. There are several possible ways of choosing what is too small for this purpose, and one of the simpler criteria is used below (see Wiggins 1972, for a detailed discussion).

To derive the resolution matrix, suppose that the observations, and therefore also the elements of the Hessian (calculated for the cost function with the real data only, without the smoothing terms), are given in p significant decimal digits. Then only eigenvalues γ_i of $\mathbf{G}'$ satisfying

$$\gamma_i/\gamma_1 > 10^{-p} \quad (12)$$

are significantly different from zero. The eigenvectors corresponding to the eigenvalues that are too small are the linear combinations of model parameters about which the data have no information. Given that k eigenvalues of $\mathbf{G}'$ are significantly different from zero, the search for the minimum point in the parameter space is performed only in the direction of the first k corresponding eigenvectors of $\mathbf{G}'$. In order for a given parameter α_{l_0} to be resolved independently of other α_l, a search should be possible in the direction of the vector δ_{l_0} (1 in the l_0 location, zeros elsewhere). In other words, α_{l_0} is resolved if δ_{l_0} can be written as a linear combination of the first k eigenvectors $\mathbf{V}_i$ of $\mathbf{G}'$. It can be shown (Wunsch 1978) that the closest one can get to δ_{l_0} is with the linear combination

$$\hat{\delta}_{l_0} = \sum_{i=1}^{k} V_{l_0 i}\mathbf{V}_i, \quad (13)$$

where $V_{l_0 i}$ is the l_0 component of the ith eigenvector $\mathbf{V}_i$. But (13) is also the l_0 column of the matrix

$$\mathbf{R} = \mathbf{V}_{L\times k}\mathbf{V}^{\mathrm{T}}_{L\times k}, \tag{14}$$

which is termed the parameter resolution matrix, where $\mathbf{V}_{L\times k}$ is a matrix whose columns are the first k eigenvectors of $\mathbf{G}'$. Having calculated the resolution matrix, we examine its columns. The diagonal elements of $\mathbf{R}$ indicate how well is each parameter resolved, while the off-diagonal elements indicate what are the other variables from which a given parameter cannot be independently resolved. (See Wiggins 1972, for more details on the use and interpretation of the resolution information.)

3) Error analysis

Suppose the cost function were simply:

$$\mathbf{J} = (\psi^0 - \hat{\psi})^{\mathrm{T}}\mathbf{A}(\psi^0 - \hat{\psi}) \tag{15}$$

where $\psi^0, \hat{\psi}$ are column vectors containing the streamfunction for the initial time and the streamfunction observations at all grid points; the following discussion can easily be extended to include the other terms appearing in the cost function given by (3). Following the standard approach to least-squares problems (Mood et al. 1974), the difference between the data and their model counterparts $\epsilon = \psi^0 - \hat{\psi}$ can be thought of as normally distributed random error with zero mean and with error covariance matrix $\mathbf{A}^{-1}$. Minimizing the cost function thus provides a maximum-likelihood estimate for the model parameters.

Let α denote the column vector of model parameters and α^* the best-fit values. Then by expanding $\mathbf{J}$ in a power series about α^*, we can see that to lowest order in α,

$$\mathbf{J} \approx \mathbf{J}_{\min} + (\alpha - \alpha^*)^{\mathrm{T}}\mathbf{G}(\alpha - \alpha^*), \tag{16}$$

where higher-order terms resulting from nonlinearities of the model have been neglected. If the neglected terms are sufficiently small, then the error in the model parameters $\Delta\alpha = \alpha - \alpha^*$ can also be assumed to be normally distributed with zero mean and with $\mathbf{C} = \mathbf{G}^{-1}$ as the error-covariance matrix (Thacker 1987).

To obtain the error covariance matrix, the Hessian $\mathbf{G}$ must be inverted. Care must be taken, however, to avoid problems of ill conditioning when some of the eigenvalues of the Hessian are not significantly different from zero, as discussed above. The inverse Hessian can be written in terms of the preconditioned Hessian as $\mathbf{C} = \mathbf{G}^{-1} = (\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{G}'^{-1}\mathbf{W}^{-1}$. Writing the preconditioned Hessian matrix as $\mathbf{G}' = \mathbf{V}\Gamma\mathbf{V}^{\mathrm{T}}$, its inverse is $\mathbf{G}'^{-1} = \mathbf{V}\Gamma^{-1}\mathbf{V}^{\mathrm{T}}$. But if there is no information about some linear combinations of parameters, the resulting very small eigenvalues will result in very large confidence intervals (terms of $\mathbf{G}^{-1}$), because of the inverse of the eigenvalues entering the calculation of $\mathbf{C}$. To avoid this problem, a generalized inverse of the Hessian must be used. With k significantly nonzero eigenvalues γ_i, $i = 1, \cdots, k$, the error covariance matrix is

$$\begin{aligned}\mathbf{C} = \mathbf{G}^{-1} &= (\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{G}'^{-1}\mathbf{W}^{-1} \\ &= (\mathbf{W}^{\mathrm{T}})^{-1}[\mathbf{V}\Gamma_k^{-1}\mathbf{V}^{\mathrm{T}}]\mathbf{W}^{-1},\end{aligned} \tag{17}$$

where Γ_k^{-1} contains $1/\gamma_i$ in the first k diagonal locations, and zeros elsewhere. However, it is important to recognize that the covariance given by (15) does not include the contribution of the combinations of parameters corresponding to the zero eigenvalues, and it is therefore an underestimate of the full error covariance matrix for those variables that are not fully resolved by the model. The covariance information must therefore be used together with the resolution information for a consistent interpretation. This is further discussed by Wiggins (1972).

3. Results

We first describe three experiments demonstrating the use of an adjoint model in estimating model parameters. For the third example, the results are analyzed using the resolution information provided by the Hessian matrix. Then, the utility of the information contained in the Hessian for experimental design is demonstrated.

Unless otherwise stated, the simulated observations used in the following experiments were of vorticity $\hat{\zeta}_{ij}$ at all grid points, and the cost function is given by (3) with $C_{ij}^{(1)} = D_{ij}^{(1)} = 0$.

First (run A), the wind-stress curl and initial vorticity fields are assumed to be unknown, while the friction parameters ϵ_b and ϵ_h are set to their correct values (those used to obtain the observations). The optimal values for the initial conditions ζ_{ij}^0 and the forcing curlτ_{ij} are calculated using the descent algorithm. Figure 2 shows the value of the cost function $\mathbf{J}$, and the distance of the initial conditions and wind curl from their correct values, as a function of iteration number. The initial guess for the parameters in this run was zero vorticity and a random (with zero mean) wind-stress curl. It is important to note that the solution for the unknown parameters was independent of the initial guess in the many experiments performed for this case.

As can be seen in Fig. 2, the descent algorithm first improves the estimate for the initial conditions, and then starts improving both the initial conditions and the wind-stress curl. Within about 175 iterations the cost function is at its minimum (zero—or in fact 10^{-12}, due to computer round-off error—as there is no observational noise). The curl is then correct to about 0.01%, as is the vorticity. Each iteration includes one time step of the forward model and one step of the adjoint model, making the calculation equivalent to about 350 forward time steps. This convergence is quite fast compared, for example, to the time needed to reach a steady solution by stepping the forward model [O(1000) time steps]. The fast convergence is partic-

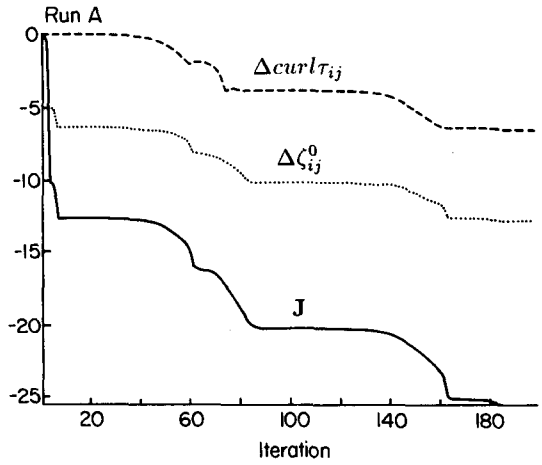

FIG. 2. Results of run A: The approach of the unknowns—wind-stress curl and vorticity initial conditions—to their correct values, as a function of iteration number. The vertical axis is the log of the least-square distance between the parameters in the optimization, and their true values. For the vorticity initial conditions this is $\log\{\sum_{ij}(\zeta^0_{ij}-\hat{\zeta}_{ij})^2\}^{1/2}$, and the curve is marked $\Delta\zeta^0_{ij}$. Similarly, the distance between the value of the wind-stress curl during the optimization and its true value is shown by the curve marked Δ curlτ_{ij}. Also shown is the value of the cost function **J**. All three curves are normalized by their initial values. Each iteration consists of one forward time step of the quasi-geostrophic model, and one backward time step of the adjoint model.

ularly encouraging, because the calculation shown in Fig. 2 involves quite a large number of unknown parameters: a 33 by 33 grid, for which there are 31×31 unknown values for the wind-stress curl and initial conditions, a total of 1922 unknowns. The whole optimization took no more than 5 cpu minutes on an IBM mainframe (model 3081D).

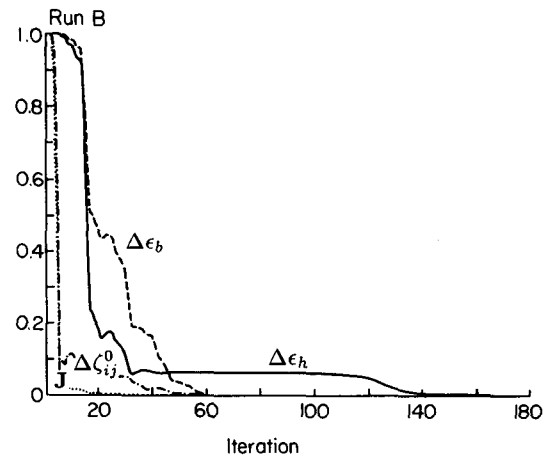

FIG. 3. Results of run B: As in Fig. 2, but this time the unknowns in the optimization are the vorticity initial conditions and the friction parameters. The curves marked $\Delta\epsilon_b$ and $\Delta\epsilon_h$, show the distance of the friction parameters from their true value as function of the iteration number. The vertical axis is their linear distance from their correct values.

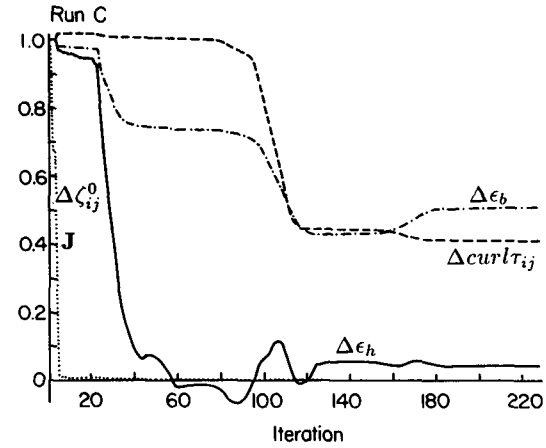

FIG. 4. Results of run C: As in Fig. 2, except that the unknowns are the vorticity initial conditions, bottom friction, horizontal friction and wind-stress curl. Note that the wind and friction parameters do not converge to their true values in this case (curves marked $\Delta\epsilon_b$, $\Delta\epsilon_h$ and Δ curlτ_{ij} do not reach zero value). Still, the initial conditions reach their true value and the cost function reaches zero value (curves for initial conditions and cost function overlap and are marked $\Delta\zeta^0_{ij}$ and **J**; both approach zero fairly quickly).

Next (run B), the wind forcing was assumed known, and set to its correct value, and the procedure was used to calculate the friction coefficients ϵ_b and ϵ_h along with the initial vorticity field. Figure 3 shows the convergence of the parameters and cost function to the optimal values, and again the convergence to the true

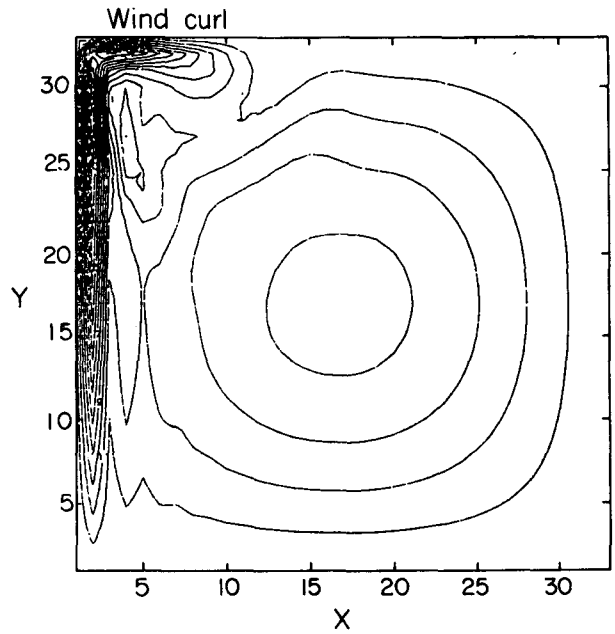

FIG. 5. Final solution for curlτ for run C shown in Fig. 4. Note the strong forcing (by curl of wind stress) in the western boundary current, balancing the dissipation there due to the too large values for the friction parameters found by the optimization.

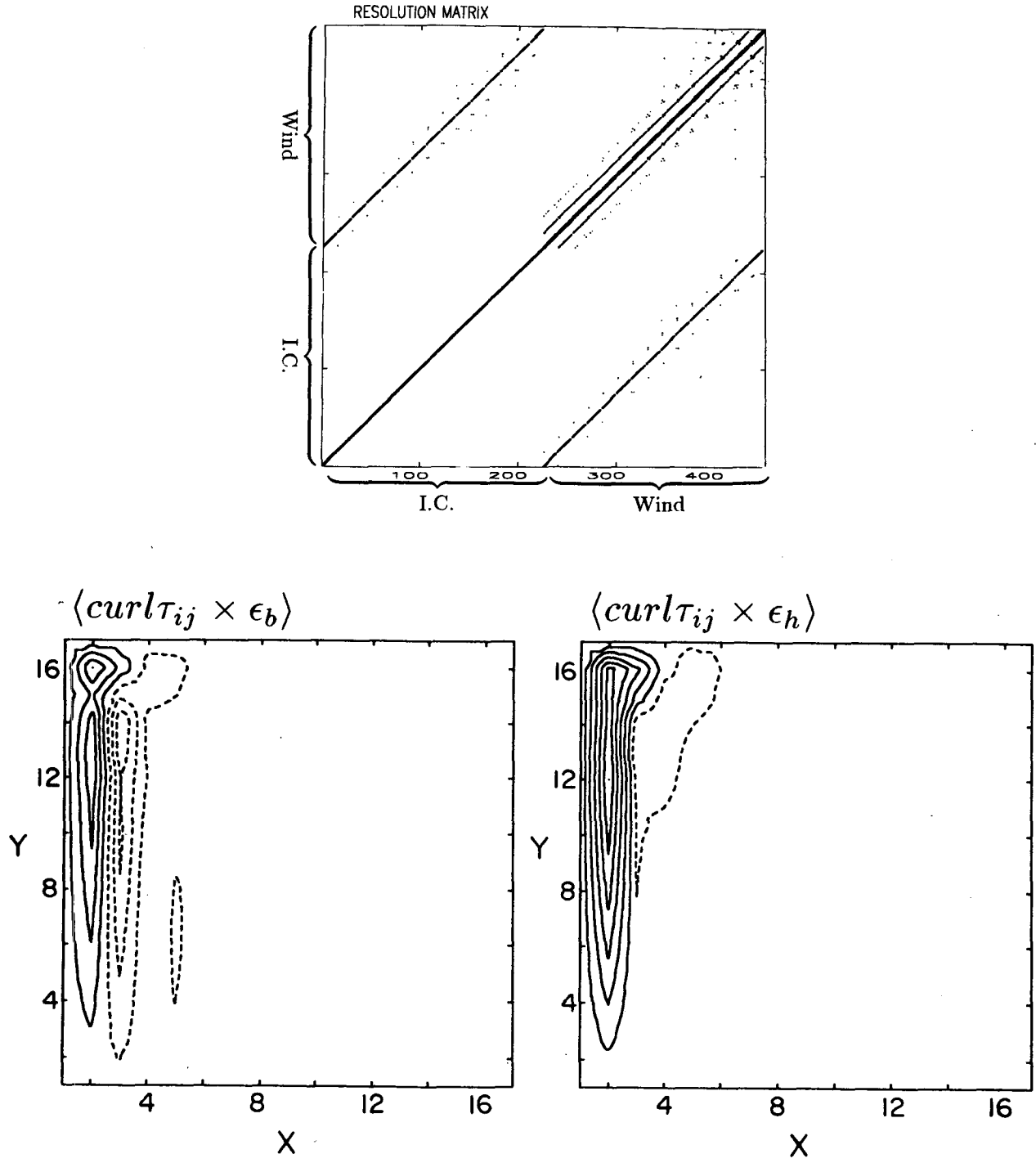

FIG. 6. (a) (upper panel) A contour plot of the resolution matrix for run C, using cutoff value of 0.001 for the eigenvalues (section 2d). Separate contours are not distinct here, but the figure gives a general impression on what elements of the resolution matrix are nonzero (black parts of the figure). The resolution information in Figs. 6 and 7 is calculated for a coarser-grid model than used in the optimization shown in Figs. 2–5 (17 by 17 grid, instead of 33 by 33 grid). The unknowns are arranged in the following order for the calculation of the resolution and error covariance matrices: $\{\text{curl}\tau_{ij}, \zeta^0_{ij}, \epsilon_b \text{ and } \epsilon_h\}$, where only $15 \times 15 = 225$ interior values for the wind-stress curl and initial conditions need to be considered. In the figure, therefore, locations 1–225 on the axes correspond to the wind-stress curl at all interior grid points, locations 226–450 to the interior initial conditions searched for the optimization, and locations 451 and 452 are for bottom and horizontal friction correspondingly. Note that the wind-stress curl and friction parameters are not fully resolved, but only in combination with the wind-stress curl at some locations. (b) (lower panels) The off-diagonal terms of the resolution matrix describing the dependence of the bottom friction and the wind-stress curl. This plot is equivalent to the first 225 locations of rows 451, and 452 in (a). It is plotted in the x–y plane to show that the solution for curlτ in the western boundary region is not independent of the friction parameters.

solution did not depend on the initial guess for the unknowns.

Finally (run C), friction parameters, wind forcing, and initial vorticity were all treated as unknowns, and were simultaneously calculated by the optimization. Poor results could be anticipated, since the number of observations is now two less than the number of unknowns. [The penalty terms multiplied by the $D_{ij}^{(2)}$ in (3), requiring the solution to be in a steady state, can be counted as bogus data at each grid point (Thacker 1988).] The descent algorithm did converge to a minimum of the cost function—in fact to $\mathbf{J} = 0$, corresponding to a steady-state solution that perfectly fits the observations—but not to the correct solution. The wind curl and friction coefficients found by the optimization were not the ones used to obtain the observations. The solution (Fig. 4) is characterized by friction coefficients larger than the true values, and wind forcing with large amplitude in the western boundary-current region. The optimal solution also varied with the initial guess for the parameters, and one such solution is shown in Fig. 5. The solution was checked by using the calculated wind and friction parameters in the forward model and stepping it a large number of time steps; it was found to be a stable steady-state solution. It seems that the larger values of the friction coefficients are balanced by vorticity input by the strong wind forcing calculated in the western boundary current region. This balance resulted in vorticity and streamfunction fields identical to the observations.

The solution for the unknown parameters should be supplemented with resolution information and error analysis, and these are given below. In examining this additional information, it will become obvious that the model cannot resolve the wind-stress curl separately from the friction parameters using these data. We had already anticipated this result, based on the fact that there were more unknowns than data, even when penalty terms were counted. Nevertheless, it is still useful to go through this additional analysis of the fit, as it illustrates how to test whether the recovered values might be wrong—even without knowing the true values.

Figure 6a shows the resolution matrix calculated at the minimum point of $\mathbf{J}$. The cutoff value for the eigenvalues used to obtain these matrices was 0.001 [$p = 3$ in (12)]. The curl of the wind stress is not fully resolved (diagonal elements of the resolution matrix smaller than one), and is not independent of the estimate obtained for the friction coefficients, as indicated by the nonzero off-diagonal terms of the resolution matrix (Fig. 6b). The error covariance matrix gives similar information on the dependence of the estimates for the wind and friction parameters. The solution for the initial conditions, on the other hand, is fairly well resolved (diagonal elements nearly one in resolution matrix), in agreement with the fact that the solution found for the vorticity is the correct one.

In order to be able to resolve both friction and wind parameters, some additional a priori information must be used. This could be, for example, a requirement that the calculated curl of the wind stress be smooth, by adding an appropriate term to the cost function. Such a term will prohibit the kind of a solution shown in Fig. 5, having large gradients of the wind-stress curl near the western boundary region. In any case, the above result of many possible solutions even for the present simple model, indicates that one has to be careful when interpreting the results of the optimization and the calculated parameters. Treating the model parameters as unknowns clearly opens many new and interesting questions not normally encountered in the more traditional approach to numerical modeling of the oceanic general circulation.

These results also raise a more general possible problem of oceanographic models that try to calculate parameters from data. In a situation where several possible solutions are consistent with the data, and none of them is unreasonable, it is obviously meaningless to try and determine the correct value of these parameters. In that case, the purpose of the modeling effort must be redefined. It may be possible and useful in such a case to try and put bounds on these unknown parameters, by looking for their minimum and maximum values that are consistent with the data, as discussed in section 2.

The information contained in the Hessian matrix can be very valuable in the process of experiment design. Given some parameters that need to be calculated (e.g., wind forcing or mixing coefficients), and limited resources, one would want to choose the observational strategy maximizing the information on the desired parameters. The choice could be, for example, between taking a hydrographic section in the western boundary current region, or in the interior, or between putting a current meter and making a hydrographic section.

As an example that is meant to be instructive rather than realistic, suppose one must decide whether to obtain streamfunction or vorticity observations in order

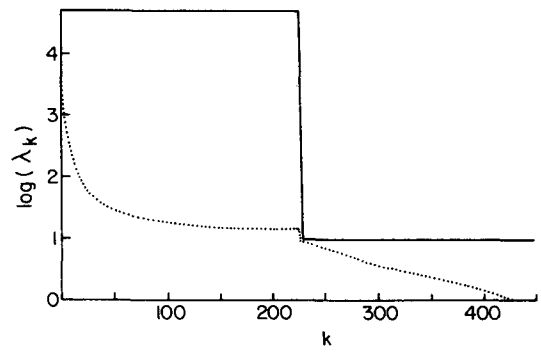

FIG. 7. The eigenvalues of the Hessian for both streamfunction (dotted line) and vorticity (full line) cost functions ($C^{(1)} = D^{(1)} = 0$ and $C^{(2)} = D^{(1)} = 0$ in (3) correspondingly), with only the wind-stress curl and initial conditions considered unknowns.

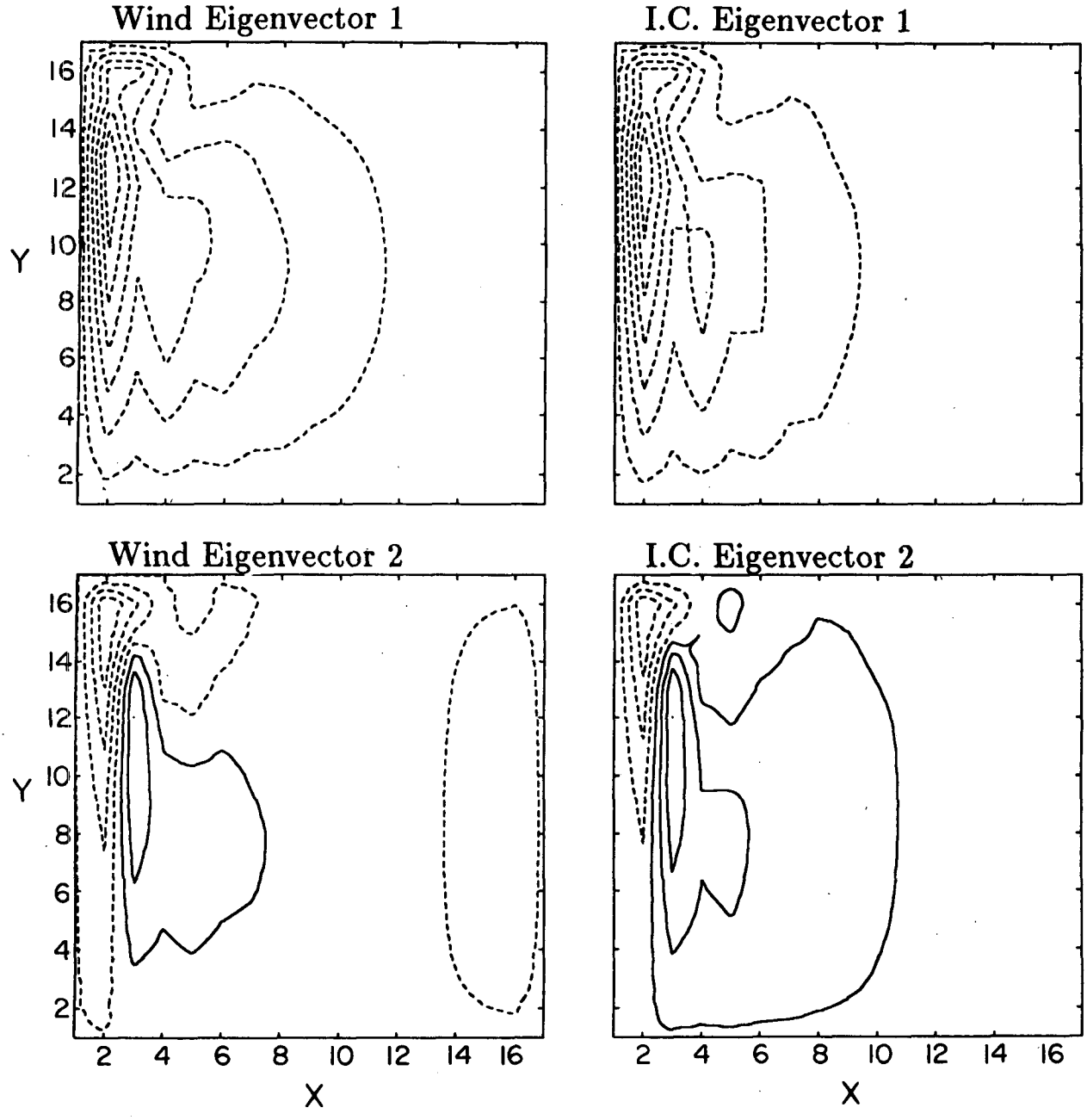

FIG. 8. First two eigenvectors of the Hessian matrix calculated for the vorticity observations, (see Fig. 7). Each eigenvector is made of wind and initial conditions parts, corresponding to locations 1–225 and 226–450 of the full eigenvectors (see caption to Fig. 6). These parts are plotted in the x–y plane.

to deduce the wind-stress curl. It is assumed (certainly not realistically), that within the experimental budget either vorticity or streamfunction can be measured at every grid point to the same relative accuracies. Two cost functions can be defined: one for streamfunction data with $C_{ij}^{(1)} = D_{ij}^{(1)} = 0$ and the other for vorticity data with $C_{ij}^{(2)} = D_{ij}^{(1)} = 0$. In both cases the steady state can be enforced using vorticity penalty terms and the initial conditions can be taken to be the initial vorticity field. We now examine which of the two datasets is more appropriate for the calculation of the wind-stress parameters.

Figure 7 shows the eigenvalues of the Hessian for the two cost functions on a logarithmic scale. [Note that a model of coarser resolution (17×17) was used for this calculation, while the (33×33) resolution was used in the actual optimizations discussed above.] The eigenvalues for the vorticity cost function (full line in

Fig. 7) are divided into two groups, and within each group all eigenvalues are very nearly of the same magnitude. Figure 8 shows the first two eigenvectors. The parts of the vectors corresponding to the wind and initial conditions parameters are plotted separately on the (x, y) plane. Examining the eigenvectors it becomes clear that the first group of eigenvectors resolves linear combinations of wind-stress curl and initial conditions without resolving either independently at any single location. This is indicated by the equal structure of the wind and initial conditions parts of the eigenvectors. The second group of eigenvectors corresponds to the separation of the wind from the initial conditions. This is seen by the similar structure, but opposite sign of the wind and initial-conditions eigenvectors for the second group (not shown in the figure). The sharp drop between the two groups of eigenvalues means that the model has difficulties in resolving the initial conditions separately from the wind-stress curl.

For the streamfunction data, the eigenvalues span a much wider range of values (dotted line in Fig. 7). This difference affects the calculation of the unknown parameters in two ways. First, the convergence of the preconditioned conjugate-gradient descent algorithm to the optimal set of parameters was much faster when the eigenvalues of the Hessian are grouped as for the vorticity observations (Gill et al. 1981). The accuracy and the resolution of the calculated parameters are also affected by the difference in range of eigenvalues for the two cases. According to the truncation criterion (12), fewer eigenvectors participate in the resolution matrix (14) for the streamfunction observations than for the vorticity observations, for a given accuracy of the observed streamfunction and vorticity. The parameters are therefore better resolved when using the vorticity observations. The smaller eigenvalues of the Hessian for the streamfunction observations also cause larger errors in the parameters, as calculated in the error covariance matrix (17). The superiority of the vorticity measurements for the calculation of the different model parameters seem to result from its being more sensitive to changes in the parameters. The streamfunction is found by twice integrating the vorticity, and its therefore less sensitive to these changes.

We may now conclude that, given the assumption of equal cost and accuracy of measurement, vorticity observations are preferable to streamfunction observations for determining the calculation of the wind-stress curl. More generally, this example shows how the information contained in the Hessian matrix may be used for the purpose of experiment design.

4. Conclusions

We have demonstrated, using a quasi-geostrophic model as an example, the use of the adjoint technique in analyzing oceanographic data with the aid of a general circulation model. The method is capable of efficiently combining a complex dynamical model, and comprehensive datasets. Unknown model parameters can be calculated, and an improved estimate for the observed fields is obtained. These are complemented with the important resolution and error information, making the method in fact a very powerful inverse procedure.

In addition, the method can be used in the process of experiment design to decide how effective particular data types are for deducing certain unknown parameters. It is also possible to predict the needed accuracy of measurements for the estimated parameters to be well resolved, and within reasonable error bounds.

The adjoint (or optimal control) method is conceptually simple, and the effort required to prepare an adjoint model for a given numerical model is not large considering its many possible uses and advantages.

As more types of data from satellites and remote sensing methods become available, and as numerical models become more and more realistic, the adjoint method can be expected to play an important role in the study of the oceanic general circulation.

Acknowledgments. Thanks to I. Feliks and I. Yavneh for helping in the development of the numerical model.

APPENDIX A

Finite-Difference Formulation of the Forward and Adjoint Models

Quasi-geostrophic model. Let ζ_{ij}^n and ψ_{ij}^n denote the vorticity and streamfunction at horizontal location (i, j), and time step n. The model is advanced in time using a diagnostic equation for the streamfunction and a prognostic equation for vorticity:

$$\nabla^2\psi_{ij}^n = \zeta_{ij}^n$$

$$(\zeta_{ij}^{n+1} - \zeta_{ij}^n)/\Delta t + (\psi_{i+1,j}^n - \psi_{i-1,j}^n)/\Delta x + RJ(\psi, \zeta)_{ij}^n = -\epsilon_b\zeta_{ij}^n + \epsilon_h\nabla^2\zeta_{ij}^n + \mathrm{curl}\tau_{ij} \quad \text{(A1)}$$

where ∇^2 and J denote the five-point Laplacian and Arakawa's nine-point conservative Jacobian (Arakawa 1966). The stream function is calculated by solving a Poisson equation, and then vorticity is advanced in time.

Adjoint model. We now write the adjoint equation in finite-difference form, with the cost function **J** unspecified. The cost function may include data distributed over N time steps and different possible choices of penalties to enforce spatial smoothness and temporal steadiness:

$$\nabla^2\mu_{ij}^n - (\lambda_{i+1,j}^{n+1} - \lambda_{i-1,j}^{n+1})/\Delta x - RJ(\lambda^{n+1}, \zeta^n)_{ij} + \frac{\partial \mathbf{J}}{\partial \zeta_{ij}^n} = 0$$

$$(\lambda_{ij}^{n} - \lambda_{ij}^{n+1})/\Delta t - RJ(\psi^n, \lambda^{n+1})_{ij} + \epsilon_b \lambda_{ij}^{n+1} - \epsilon_h \nabla^2 \lambda_{ij}^{n+1} - \mu_{ij}^{n} + \frac{\partial \mathbf{J}}{\partial \zeta_{ij}^{n}} = 0. \quad \text{(A2)}$$

The adjoint variable λ is stepped (backward) in time, while μ is calculated by solving a Poisson equation at every step. Note that ∇^2 and J again represent the five-point Laplacian and Arakawa's Jacobian. Also note that the partial derivatives of $\mathbf{J}$ are evaluated with $\mathbf{J}$ considered to be a function of *all* model variables. For the cost function (3), involving only time levels $n = 0$ and $n = 1$, the equations for λ_{ij}^{1}, μ_{ij}^{1} and μ_{ij}^{0} are

$$\nabla^2 \mu_{ij}^{1} = -2D_{ij}^{(1)}(\psi_{ij}^{1} - \psi_{ij}^{0})$$

$$\lambda_{ij}^{1}/\Delta t = \mu_{ij}^{1} - 2D_{ij}^{(2)}(\zeta_{ij}^{1} - \zeta_{ij}^{0}).$$

$$\nabla^2 \mu_{ij}^{0} = (\lambda_{i+1,j}^{1} - \lambda_{i-1,j}^{1})/\Delta x + RJ(\lambda^1, \zeta^0)_{ij} - 2[C_{ij}^{(1)}(\psi_{ij}^{0} - \hat{\psi}_{ij}) + D_{ij}^{(1)}(\psi_{ij}^{0} - \psi_{ij}^{1})]. \quad \text{(A3)}$$

There are no λ_{ij}^{0} multiplying any equation in the definition of the Lagrange function, but it is convenient to introduce such variables, since λ_{ij}^{0} can be recognized to be the negative of the components of the gradient of the cost function in the directions associated with the initial conditions; they can be calculated by stepping the adjoint equation for λ from $n = 1$ to $n = 0$.

$$\lambda_{ij}^{0}/\Delta t = \lambda_{ij}^{1}/\Delta t + RJ(\psi^0, \lambda^1)_{ij} - \epsilon_b \lambda_{ij}^{1} + \epsilon_h \nabla^2 \lambda_{ij}^{1} + \mu_{ij}^{0} - 2[C_{ij}^{(2)}(\zeta_{ij}^{0} - \hat{\zeta}_{ij}) + D_{ij}^{(2)}(\zeta_{ij}^{0} - \zeta_{ij}^{1})]. \quad \text{(A4)}$$

APPENDIX B

Derivation of the Adjoint Equations in Continuous Form

We now derive the continuous form of the adjoint equation from the continuous quasi-geostrophic equation (1) and the continuous form of the cost function (3), using the calculus of variations (Courant and Hilbert 1953).

As in the finite-difference form, we wish to minimize the cost function measuring the distance between observations and model results, subject to the dynamic equations as constraints. For the continuous derivation, the data are assumed to be sufficiently continuous and differentiable, even though this restriction is not needed in the discrete case. The continuous cost function is (omitting the terms multiplying the $D^{(k)}$, which can easily be added to the derivation)

$$\mathbf{J}(\text{curl}\tau, \epsilon_b, \epsilon_h) = \int_{x=0}^{L_x} dx \int_{y=0}^{L_y} dy \int_{t=0}^{T} dt \times [C^{(1)}(x, y, t)(\psi(x, y, t) - \hat{\psi}(x, y))^2 + C^{(2)}(x, y, t)(\zeta(x, y, t) - \hat{\zeta}(x, y))^2] \quad \text{(B1)}$$

where T is the time period over which the model is run at each iteration. Note that time-independent data are compared with the model at all times; for consistency with Eq. (3), the weight functions $C^{(i)}$ should be different from zero only near the initial time $t = 0$. The derivation is somewhat simpler if we write the quasi-geostrophic equation as two equations

$$\zeta_t + \psi_x + RJ(\psi, \zeta) = -\epsilon_b \zeta + \epsilon_h \nabla^2 \zeta + \text{curl}\tau$$

$$\nabla^2 \psi = \zeta. \quad \text{(B2)}$$

Forming a Lagrange function by adding the constraints multiplied by the Lagrange multipliers, we have

$$\mathbf{L}(\text{curl}\tau, \epsilon_b, \epsilon_h, \psi, \zeta, \lambda, \mu) = \mathbf{J}(\text{curl}\tau, \epsilon_b, \epsilon_h) + \int_{x=0}^{L_x} dx \int_{y=0}^{L_y} dy \int_{t=0}^{T} dt \{\lambda(x, y, t)[\zeta_t + \psi_x + RJ(\psi, \zeta) + \epsilon_b \zeta - \epsilon_h \nabla^2 \zeta - \text{curl}\tau] + \mu(x, y, t)[\nabla^2 \psi - \zeta]\}. \quad \text{(B3)}$$

Actually, the boundary conditions should also be introduced as constraints in order to deduce the boundary conditions for the adjoint equations, but since the procedure is similar to that for the interior equations, those details are omitted to simplify the discussion.

At the minimum of $\mathbf{J}$, $\mathbf{L}$ has a stationary point. Its first variation with respect to all of its arguments must vanish there. The first variation of the Lagrange multipliers λ and μ gives the original model equations. The variation with respect to ζ is

$$\delta \mathbf{L} = \int_{x=0}^{L_x} dx \int_{y=0}^{L_y} dy \int_{t=0}^{T} dt \{2C^{(2)}(\zeta - \hat{\zeta})\delta\zeta + \lambda[\delta\zeta_t + \psi_x + RJ(\psi, \delta\zeta) + \epsilon_b \delta\zeta - \epsilon_h \nabla^2 \delta\zeta] - \mu(x, y, t)\delta\zeta\} = 0.$$

After integration by parts, since the boundary conditions for ψ and ζ cause the boundary terms to vanish,

$$\delta \mathbf{L} = \int_{x=0}^{L_x} dx \int_{y=0}^{L_y} dy \int_{t=0}^{T} dt \delta\zeta(x, y, t) \times \{2C^{(2)}(\zeta - \hat{\zeta}) - \lambda_t - RJ(\psi, \lambda) + \epsilon_b \lambda - \epsilon_h \nabla^2 \lambda - \mu\} = 0, \quad \text{(B4)}$$

Similarly, the variation with respect to the streamfunction gives after integration by parts

$$\delta \mathbf{L} = \int_{x=0}^{L_x} dx \int_{y=0}^{L_y} dy \int_{t=0}^{T} dt \delta\psi(x, \quad t) \times \{2C^{(1)}(\psi - \hat{\psi}) - \lambda_x - RJ(\lambda, \zeta) + \nabla^2 \mu\} = 0, \quad \text{(B5)}$$

The coefficients of $\delta\psi$ in (B5) and $\delta\zeta$ in (B4) must vanish, which gives us two equations for λ and μ. Substituting one in the other, we obtain the adjoint (5) to Eq. (1).

One disadvantage of this continuous derivation is that it does not prescribe the discrete form of the differential operators that appear in the adjoint equations.

It is important that they be the adjoints of the finite-difference operators that are used in the model. A careful derivation based on the discrete equations shows that centered spatial differences and Arakawa's Jacobian should also be used in the adjoint model, whereas the model's forward time step becomes a backward time step in the adjoint model. Also, because the solution of the Poisson equation (A1) in the forward model may be an approximate solution, it is important that the Poisson solver in the adjoint model (A2) be the adjoint of the approximate Poisson solution of the forward run. This is achieved by introducing Lagrange multipliers at each step of the algorithm, just as they are introduced for each time step of the prognostic equation. However, rather than doing this, we took the alternative of forcing the Poisson solver to converge to machine accuracy for all Poisson solutions, so that the question of finding the appropriate adjoint of an unconverged approximation could be avoided.

REFERENCES

Arakawa, A., 1966: Computational design for long term numerical integration of the equations of fluid motion: Two dimensional incompressible flow. Part 1. *J. Comput. Phys.*, **1**, 119–143.

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

Cacuci, D. G., 1981: A sensitivity theory for nonlinear systems. I. Nonlinear functional analysis approach. *J. Math. Phys.*, **22**, 2794–2802.

Courant, R., and D. Hilbert, 1953: *Methods of Mathematical physics, Vol. 1.* Interscience, 501 pp.

Gill, P. E., W. Murray and M. H. Wright, 1981: *Practical Optimization.* Springer-Verlag, 377 pp.

Hall, M. C. G., 1986: Application of adjoint sensitivity theory to an atmospheric general circulation model. *J. Atmos. Sci.*, **43**, 2644–2651.

——, and D. G. Cacuci, 1983: Physical interpretation of the adjoint functions for sensitivity analysis of atmospheric models. *J. Atmos. Sci.*, **40**, 2537–2546.

Hasdorff, L., 1976: *Gradient Optimization and Nonlinear Control.* Wiley and Sons, 264 pp.

Lanczos, C., 1961: *Linear Differential Operators.* D. Van Nostrand, 564 pp.

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

Mood, A. M., F. A. Graybill, and D. C. Boes, 1974: *Introduction to the Theory of Statistics.* McGraw-Hill, 564 pp.

Numerical Algorithms Group, 1984: Fortran Library Manual, Mark II, 6 vols.

Olbers, D. J., M. Wenzel and J. Willebrand, 1985: The inference of North Atlantic circulation patterns from climatological hydrographic data. *Rev. Geophys.*, **23**, 313–356.

Pedlosky, J., 1979: *Geophysical Fluid Dynamics.* Springer-Verlag, 624 pp.

Schröter, J., and C. Wunsch, 1986: Solution of nonlinear difference ocean models by optimization methods with sensitivity and observational strategy analysis. *J. Phys. Oceanogr.*, **16**, 1855–1874.

Thacker, W. C., 1987: Three lectures on fitting Numerical models to observations. *External report GKSS 87/E/65.* GKSS-Forschungszentrum Geesthacht GmbH, Geesthacht, Federal Republic of Germany, 64 pp.

——, 1988: Fitting models to inadequate data by enforcing spatial and temporal smoothness. *J. Geophys. Res.*, **93**, 10 655–10 665.

——, and R. B. Long, 1988: Fitting dynamics to data. *J. Geophys. Res.* **93**(C2), 1227–1240.

Veronis, G., 1966: Wind driven ocean circulation—Part 2. numerical solutions of the nonlinear problem. *Deep-Sea Res.*, **13**, 31–55.

Wiggins, R. A., 1972: The general linear inverse problem: Implication to surface waves and free oscillations on earth structure. *Rev. Geophys.*, **10**, 251–285.

Wunsch, C., 1978: The general circulation of the North Atlantic west of 50°W determined from inverse methods. *Rev. Geophys.*, **16**, 583–620.

——, 1984: An eclectic Atlantic Ocean circulation model, part 1: The meridional flux of heat. *J. Phys. Oceanogr.*, **14**, 1712–1733.

——, 1988: Transient tracers as a problem in control theory. *J. Geophys. Rev.*, **93**(C7), 8099–8110.

Ypma, T. J., 1987: Efficient estimation of sparse Jacobian matrices by differences. *J. Comp. Appl. Math.*, **18**, 17–28.