Available online at www.sciencedirect.com

ScienceDirect

Ocean Modelling 16 (2007) 264–276

Ocean Modelling

www.elsevier.com/locate/ocemod

Data assimilation with inequality constraints

W.C. Thacker

Atlantic Oceanographic and Meteorological Laboratory, 4301 Rickenbacker Causeway, Miami, FL 33149, United States

Received 26 September 2006; received in revised form 2 November 2006; accepted 3 November 2006
Available online 1 December 2006

Abstract

If values of variables in a numerical model are limited to specified ranges, these restrictions should be enforced when data are assimilated. The simplest option is to assimilate without regard for constraints and then to correct any violations without worrying about additional corrections implied by correlated errors. This paper addresses the incorporation of inequality constraints into the standard variational framework of optimal interpolation with emphasis on our limited knowledge of the underlying probability distributions. Simple examples involving only two or three variables are used to illustrate graphically how active constraints can be treated as error-free data when background errors obey a truncated multi-normal distribution. Using Lagrange multipliers, the formalism is expanded to encompass the active constraints. Two algorithms are presented, both relying on a solution ignoring the inequality constraints to discover violations to be enforced. While explicitly enforcing a subset can, via correlations, correct the others, pragmatism based on our poor knowledge of the underlying probability distributions suggests the expedient of enforcing them all explicitly to avoid the computationally expensive task of determining the minimum active set. If additional violations are encountered with these solutions, the process can be repeated. Simple examples are used to illustrate the algorithms and to examine the nature of the corrections implied by correlated errors.
Published by Elsevier Ltd.

Keywords: Oceanic data assimilation; Estimation theory; HYCOM; Inequality constraint

1. Introduction

This examination of data assimilation with inequality constraints has been motivated by the problem of assimilating data into HYCOM[1] (Thacker and Esenkov, 2002; Thacker et al., 2004), a numerical model of oceanic circulation that is distinguished by its hybrid vertical coordinate (Bleck, 2002). Within and below the thermocline, a model layer is density-like, while it is pressure-like in the surface mixed layer. Because the nature of the layer can vary spatially and temporally during a simulation, a data-assimilation scheme must be able to correct both its thickness and its density. When correcting density there is generally no problem in maintaining positivity, as the density of sea water is sufficiently greater than zero. On the other hand, when optimal interpolation (Gandin, 1963; Bretherton et al., 1976; Daley, 1991; Behringer et al., 1998; Carton

E-mail address: carlisle.thacker@noaa.gov

[1] HYCOM stands for HYbrid-Coordinate Ocean Model.

1463-5003/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.ocemod.2006.11.001

et al., 2000) is used to correct layer thickness, the result can be less than the model's definition allows. In such a case, it is simple enough to increase the layer's thickness to its minimum allowed value, but there is the question of whether values of other variables should also be adjusted. The purpose of this paper is to point out that, yes, in principle they should be adjusted and to provide a framework for evaluating the practical importance of such adjustments. Because inequality constraints dramatically increase the computational burden of finding an exact solution for optimal assimilation, algorithms for less expensive approximate solutions are presented.

While this study has been motivated by HYCOM's numerical formulation, inequality constraints are quite common and can be dictated by physical requirements. For example, freezing and evaporation limit sea water's temperature. And, by definition, concentrations of chemical or of biological species are necessarily positive. Even without data assimilation, such constraints can be numerically challenging (Smolarkiewicz, 1984; Gnanadesikan, 1999), and they should not be ignored when assimilating data. Although the context for this discussion is provided by the numerically based constraints on layer thickness, the analysis is equally suited to physically based inequality constraints.

Constrained assimilation is approached here via the variational framework supporting optimal interpolation. Corrections to the model state are based on assumptions about the statistical distributions of its errors and those of the observations that are assimilated; to the extent that these distributions are correct, the resulting estimate is *optimal*. The same is true when a Kalman filter (Kalman, 1960; Gelb, 1974; Evensen, 1994; Cohn, 1997) is used to assimilate data, the only difference being an attempt to account for the dynamical evolution of the state errors. And as the so-called adjoint method (Le Dimet and Talagrand, 1986; Thacker and Long, 1988; Marotzke et al., 1999) shares the same variational foundations as optimal interpolation and Kalman filtering, this discussion is also relevant to that method. This variational framework is used here to analyze the impact of inequality constraints. The mathematical problem to be solved can be recognized to be one of nonlinear programming for which methods of solution exist (Gill et al., 1981). However, the focus here is not on solving the nonlinear programming problem *per se*, as such solutions can be expensive, but on understanding the cost versus benefit of such a solution relative to the simple expedient of a post-assimilation correction to enforce the constraints and on finding a reasonably inexpensive approximate solution.

Error distributions are at the heart of optimal methods for assimilating data (Thacker, 1989). If layer thicknesses were always prescribed and not allowed to evolve during a simulation, then there would be no thickness error and uncertainties would manifest in the layer's density. Similarly, if layers were always density-like, there would be no question about the value of density, and errors would manifest in thickness. For a hybrid model where this uncertainty is spread over both density and thickness, an accurate description of uncertainties is especially challenging.[2] Following Thacker and Esenkov (2002) the uncertainties of a layer's thickness are assumed to be described by a multi-normal distribution in the absence of range constraints. As this distribution permits layers to be arbitrarily thin and even to have negative thickness, range constraints clearly demand a different characterization of uncertainties. The approach taken here is to truncate the multi-normal distribution so that its shape is retained for allowed values and forbidden values will have a probability of zero. By avoiding the more complex but less pressing issue of what might be a more appropriate error model, attention is focused on the consequences of having range constraints.

Section 2 discusses the standard variational formalism of data assimilation without constraints, emphasizing its statistical interpretation and the difficulty in knowing the correct error statistics. Simple cases are examined to clarify the formalism and to motivate the algorithms. Inequality constraints are introduced into the formalism in Section 3, and algorithms for enforcing them are derived. Section 4 provides simple computational examples taken from the context of assimilating data into a hybrid-coordinate model, focusing on four adjacent grid cells for one near-surface layer and the consequences of assimilating a single observation. These examples illustrate how variables that are not directly involved in any constraint violations can be impacted, when their errors are correlated with those of a variable that is directly involved. Section 5 concludes that,

[2] Thacker and Esenkov (2002) have suggested a reciprocal relationship between the uncertainty of a layer's density and that of its thickness: corrections to density are restricted when the layer's nature is thought to be isopycnic and those to thickness when isobaric, while the restrictions are relaxed when its nature is in transition. The task of specifying error distributions that reflect the changing nature of the vertical coordinate is an issue worth further exploration.

given our limited knowledge of error statistics and our need for computational efficiency, the simple expedient of a post-assimilative enforcement of violated constraints is not unreasonable and that, if computational resources allow, the algorithms presented here offer a practical approach to computing second-order effects.

2. Formalism

Before addressing modifications to encompass range constraints, it is useful to review the standard variational formalism that forms the basis for a wide variety of methods used to assimilate data into numerical models (Lorenc, 1986; Evensen, 1994). Its objective is to reach a compromise **x** between some background estimate **b** of the model state and additional information provided by data **d**. For this discussion, each component of the vector **x** can be regarded as a possible value for the thickness of one of the model's layers for a particular grid cell, and those of **b** as estimates for these thickness produced by the model before data have been assimilated. The data vector **d** contains values of layer thickness inferred[3] from data for the grid cells that have been observed, their model counterparts[4] being **Hx**. The compromise that is sought should have **x** close to **b** and **Hx** close to **d**. Exactly how close should depend on the relative validity of the two sources of information. If uncertainties about **b** and **d** are characterized by multivariate normal distributions with covariance matrices **B** and **D**, respectively, the compromise estimate can be defined by the minimum of the objective function:

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{b})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b}) + \frac{1}{2}(\mathbf{Hx}-\mathbf{d})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{Hx}-\mathbf{d}). \tag{1}$$

The function J increases quadratically as **x** deviates from the background estimate and as **Hx** deviates from the data with the rate of increase governed by the levels of uncertainty. The vector **x** that minimizes J can be written as a correction to the background **b**:

$$\mathbf{x} = \mathbf{b} + \mathbf{BH}^{\mathrm{T}}(\mathbf{HBH}^{\mathrm{T}} + \mathbf{D})^{-1}(\mathbf{d} - \mathbf{Hb}). \tag{2}$$

Having good estimates of the covariance matrices **B** and **D**, which account for the uncertainties in the **b** and **d**, is important for obtaining the optimal compromise, but in practice these matrices are not know with great precision. Moreover, when dealing with thickness of model layers that might at times be isobaric and at other times isopycnic, errors are particularly difficult to quantify. So the compromise **x** should be understood as reflecting this imprecision, even though it is generally referred to as being *optimal*. Assuming the error statistics are in fact correct, i.e., ignoring their imprecision, the distribution of the errors of optimal estimate **x** is multinormal. In terms of the gain matrix:

$$\mathbf{K} = \mathbf{BH}^{\mathrm{T}}(\mathbf{HBH}^{\mathrm{T}} + \mathbf{D})^{-1}, \tag{3}$$

which converts model-data differences into corrections to the model state, the covariance matrix characterizing the uncertainty of the optimal estimate is:

$$\mathbf{X} = (\mathbf{I} - \mathbf{KH})\mathbf{B}. \tag{4}$$

Updated uncertainty is less than that of the background estimate, because of the information provided by the assimilated data.

Before addressing how inequality constraints change the solution, it is instructive to examine the formalism for the simplest possible situations. First, consider the case of a model state having only one variable; in this case the vectors **x**, **b**, and **d** have only one element, as do the matrices **B**, **D**, and **H**. For sake of discussion we can think of the one variable as being the pressure thickness of one of the model's layers in one of the grid cells. If b_1 and d_1 represent background and observation for this thickness and if $\sigma^2_{b_1}$ and $\sigma^2_{d_1}$ characterize their respective uncertainties, then the optimal estimate for this one-variable case is:

[3] A procedure for inferring layer thickness from hydrographic data is discussed by Thacker and Esenkov (2002).

[4] The matrix **H** reduces the vector **x** of model variables to the smaller vector **Hx** of model counterparts of the data. Within the current context, **H** can be obtained from the identity matrix **I** by removing rows corresponding to unobserved cells.

$$x_1 = \left(\frac{b_1}{\sigma_{b_1}^2} + \frac{d_1}{\sigma_{d_1}^2}\right) \Bigg/ \left(\frac{1}{\sigma_{b_1}^2} + \frac{1}{\sigma_{d_1}^2}\right). \tag{5}$$

The compromise is a weighted average of the background and data estimates with weights proportional to the reciprocals of the respective variances, so the optimal estimate is dominated by the more trustworthy contribution. Because b_1 and d_1 should both be greater than the minimum allowed value, it is easy to see that the optimal compromise x_1 is too, so the constraint is no problem for this one-variable case.

Now consider the case of a model state with two variables, only one of which has been observed. The vectors **x** and **b** have two elements, **B** is a 2 × 2 matrix with off-diagonal elements that allow information from the single observation to influence both variables. **H** is a 2 × 1 matrix with elements 1 and 0, while **d** and **D** have only one element. Think of the two variables as being pressure thicknesses of one of the model's layers for two neighboring cells, and think of the data as an observation of the thickness for cell-1. The presence of cell-2 does not alter the optimal estimate for cell-1; it is still given by (5), but it can be written in update form:

$$x_1 = b_1 + \frac{\sigma_{b_1}^2}{\sigma_{d_1}^2 + \sigma_{b_1}^2}(d_1 - b_1). \tag{6}$$

But what does the new measurement say about x_2?

$$x_2 = b_2 + \frac{\mathrm{cov}(b_2, b_1)}{\sigma_{d_1}^2 + \sigma_{b_1}^2}(d_1 - b_1). \tag{7}$$

If $\sigma_{d_1}^2$ did not appear in the denominator of the expression (7) for updating x_2, then it would be the usual linear–regression formula[5] for predicting the thickness for cell-2 from observations for cell-1; the presence of $\sigma_{d_1}^2$ accounts for the errors in the observations (Draper and Smith, 1981). Combining (6) and (7) shows that x_1, the optimal estimate of the thickness for cell-1, can be used as "data without error" for a regression estimate of the thickness for cell-2:

$$x_2 = b_2 + \frac{\mathrm{cov}(b_2, b_1)}{\sigma_{b_1}^2}(x_1 - b_1). \tag{8}$$

Fig. 1 illustrates Eq. (8) geometrically. The ellipses are contours of a Gaussian probability-density function describing the uncertainty of a background estimate having a value of 2 for cell-1's layer thickness and 1 for cell-2's. The elliptical nature of the contours is due to the between-cell correlations, which require updating cell-2 even when only cell-1 has been observed. If hydrographic data require a greater value for cell-1's layer thickness and Eq. (6) produces an updated value of 2.5, then the maximum of the probability density along the vertical line, which is marked by a circle, determines the updated value for cell-2. If the update for cell-1 had been different, so would the update for cell-2; the diagonal line indicating this relationship can be regarded as a regression line.

Fig. 2 illustrates how constraints can change things. Shading indicates forbidden values for which the probability density should be zero, so the elliptical contours are for a probability-density function that has a truncated range. The maximum probability density is at coordinates (2, 1), indicating that the background estimate for cell-2's layer thickness is at the limit of its range, and the orientation of the contours indicate a positive between-cell correlation. So any observation that updates cell-1's thickness to a higher value will cause cell-2's layer thickness to increase beyond its minimum, but those that decrease thickness for cell-1 will cause cell-2's to violate its constraint. These two cases correspond to the two vertical lines, one to the right and the other to the left of the center of the elliptical contours; the intersection of the left-hand line with the regression line is in the forbidden zone. The presence of inequality constraints changes the probabilistic interpretation from one based on multi-normal distributions to one based on truncated multi-normals, and constraint enforcement puts the optimal solution at the point on the left vertical line where the probability density is maximum; this point is on the boundary of the feasible region.

[5] The covariance cov(b_2,b_1) and variance $\sigma_{b_1}^2$ together with the background values b_1 and b_2 might be recast as slope and intercept coefficients obtained by fitting to data.

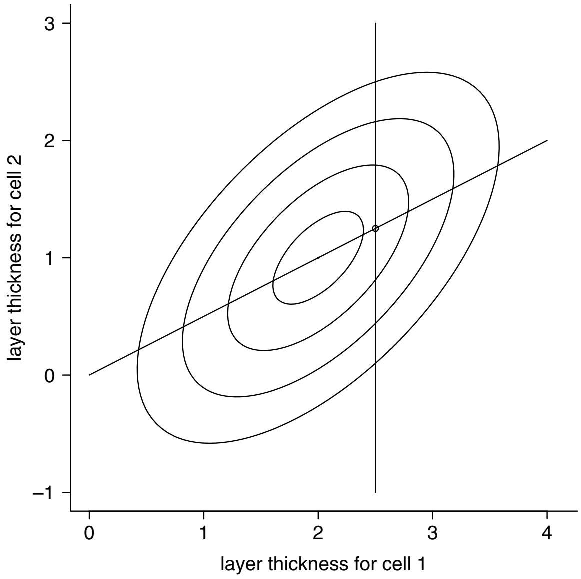

Fig. 1. Elliptical contours indicate levels of a probability-density function describing the uncertainty of a background estimate for two variables. Its maximum at the point (2, 1) corresponds to the background values for layer thickness for two cells. If cell-1 thickness after assimilation is increased to 2.5, then cell-2's is given by the maximum of the probability density along the vertical line, which is marked by the circle. Diagonal line indicates the most likely thickness for cell-2 for each possible corrected value for cell-1.

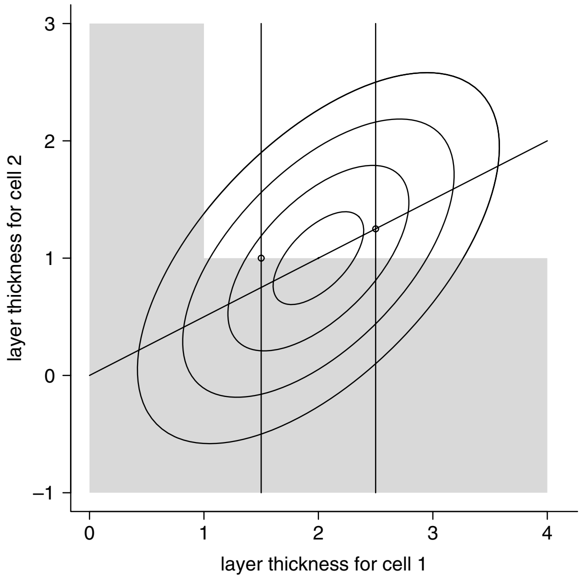

Fig. 2. Elliptical contours indicate levels of the background probability-density function as in Fig. 1. Shading indicates region where inequality constraints are violated. Vertical lines indicate two possible updated values for cell-2, and their intersections with the diagonal regression line indicates the updated value for cell-2 when constraints are ignored. Circles mark the updated values when constraints are enforced.

Now consider a case involving three cells, but new data for only one. Again, the new estimate for the observed cell is given by (5), independent of the existence of the other two cells. And the update for cell-2 is given by (8), in spite of the presence of cell-3. The expression for the update for cell-3 is obtained by replacing subscript 2 with subscript 3 in Eq. (8). Fig. 3 illustrates the solution by showing the background-error

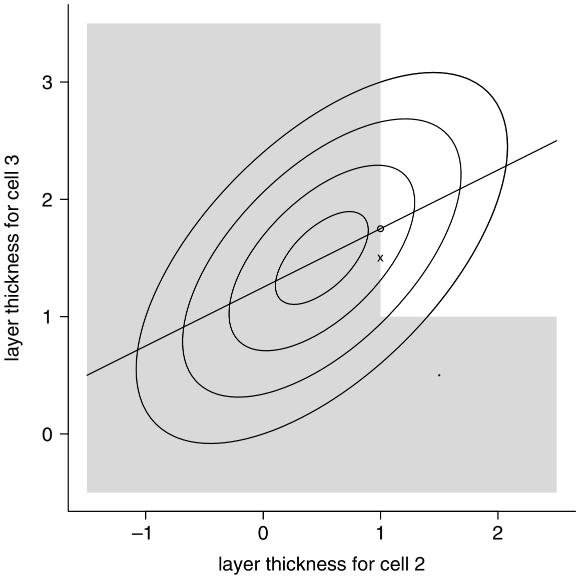

Fig. 3. Elliptical contours indicate levels of a three-variable background probability-density function within the plane determined by the updated value for cell-1. The maximum lies in the forbidden zone where cell-2's layer thickness is less than allowed. The circle marks the constrained maximum, while the cross marks the estimate obtained by simply correcting cell-2's thickness without regard for its effect on cell-3.

probability-density function in the plane determined by the updated value of the layer thickness for cell-1. In the absence of inequality constraints, the maximum of the probability density in this plane determines the updated values for thickness of the unobserved cells. The maximum is shown to occur at coordinates $(0.5, 1.5)$, which is within the forbidden zone; the problem is not with cell-3's updated thickness, which is 1.5 times its minimum allowed value, but with cell-2's being half its minimum allowed value. Clearly, thickness for cell-2 must be increased to unity. But what about cell-3? Positive between-cell correlations require cell-3's thickness to be increased along with cell-2's. Its optimal value is given by the maximum along the constraint line, which occurs at the value of 2.0. Thus, *correlated background errors can require changes to variables other than those that overtly violate inequality constraints.*

Similarly, when several variables violate constraints, the optimal solution might not require that each be enforced explicitly. This is illustrated in Fig. 4, which differs from Fig. 3 both in the location of the maximum of the probability density and in the degree to which the variables are correlated. After assimilating data for cell-1, the updates without regard for the constraints has both cell-2 and cell-3 unacceptably thin. The optimal solution is given by the maximum probability density on the boundary of the allowed region, which occurs at the point where cell-2's thickness is its minimum allowed value but cell-3's is larger than its minimum.

These examples have presumed that the optimum estimate that combines a background estimate and new data is defined by the minimum of the quadratic objective function J of (1) within the feasible region. From the statistical perspective this implies a truncated multi-normal distribution of background errors. This raises the question of whether this is appropriate statistical distribution to use. In the absence of constraints the multi-normal distribution was convenient for several reasons. First, it requires a minimum of parameters to specify, namely the variances and covariances that determine a multi-dimensional ellipsoidal region defining the spread of possibilities. However, these numbers are generally quite poorly known. A second reason is that such Gaussian distributions lead to nice, closed-form expressions like those of Eqs. (2)–(4), which can be easily implemented. Alternatively, J can be minimized efficiently using algorithms like conjugate-gradient descent (Gill et al., 1981). Such computational issues dominate when little is known about the true nature of the error distribution, hardly enough even to quantify its spread. While inequality constraints do provide information about where the probability should vanish, they say nothing about the form of the distribution within the feasible region. Again, the lack of precise knowledge about the probabilities dictates that computational

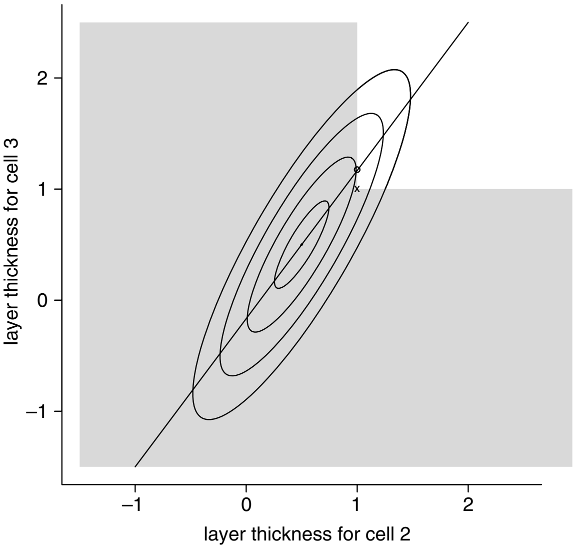

Fig. 4. Contours indicate stronger between-cell correlations than those of Fig. 3 and their center corresponds to constraint violations for both cell-2 and cell-3. The circle marks the maximum of the probability density along the boundary of the feasible region at a point where cell-2's layer thickness is the minimum possible but cell-3's is greater than minimum. The cross marks the point where both are at their minima.

convenience should be given highest consideration. It is important to keep in mind that changes to the assumed error distribution, i.e., changes to the definition of J will lead to different solutions x. Which among these is *best* cannot be determined without better knowledge how to characterize the uncertainty of the error distributions. Thus, a practical approach to enforcing inequality constraints is to avoid computationally expensive algorithms that provide a precise minimum of a poorly known J in favor of cheaper alternatives that get close to its minimum.

3. Algorithms

The preceding discussion suggests that inequality constraints can be ignored if they are not violated and can be treated like *equality* constraints when they need to be enforced. However, it did not address how to determine which of the inequality constraints should be converted to equality constraints and which might be ignored. A set of candidates for such conversion is easily obtained by first assimilating without range constraints and then checking which have been violated. As Fig. 4 illustrates, not all of these need be converted to equality constraints, because correlations can guarantee the enforcement of the remainder. If our ability to characterize uncertainties were less limited so that the definition of the objective function were less arbitrary, knowing exactly which to convert to equality constraints might be more important. In any case it can be set aside for now, as the algorithms developed here can be used to explore this issue later.

Undetermined Lagrange multipliers (Thacker and Long, 1988) offer a simple way to incorporate equality constraints into the variational formalism. If P is a matrix similar to H with the job of building expressions for the equality constraints rather than model counterparts of the observations, then the constraints can be written:

$$\mathsf{P}\mathbf{x} = \mathbf{c}, \tag{9}$$

where the vector $\mathbf{c}$ contains the constraint values, i.e., the minimum thicknesses which need to be enforced. The solution we seek is the minimum of J in (1) under the restriction that (9) is satisfied. The constrained minimum of J corresponds to a stationary point of the Lagrange function:

$$L(\mathbf{x}, \mathbf{y}) = J(\mathbf{x}) - (\mathsf{P}\mathbf{x} - \mathbf{c})^{\mathrm{T}}\mathbf{y}. \tag{10}$$

The vector $\mathbf{y}$ contains the Lagrange multipliers that enforce the constraints; their values must be determined as part of the solution.

The solution for the constrained minimum can best be seen when the objective function J is written in terms of the unconstrained minimum $\mathbf{x}_0$:

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\mathrm{T}}\mathbf{X}_0^{-1}(\mathbf{x} - \mathbf{x}_0) + J(\mathbf{x}_0), \tag{11}$$

where $\mathbf{X}_0 = \mathbf{B} - \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{D})^{-1}\mathbf{H}\mathbf{B}$ is the posterior covariance matrix. Stationarity of L with respect to variations of $\mathbf{x}$ gives an expression for the constrained solution as a correction to the unconstrained solution: $\mathbf{x} = \mathbf{x}_0 - \mathbf{X}_0\mathbf{P}^{\mathrm{T}}\mathbf{y}$. Substituting this result into the constraint Eq. (9) gives an expression for evaluating the multipliers: $\mathbf{y} = (\mathbf{P}\mathbf{X}_0\mathbf{P}^{\mathrm{T}})^{-1}(\mathbf{c} - \mathbf{P}\mathbf{x}_0)$. Thus, the constrained solution is:

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{X}_0\mathbf{P}^{\mathrm{T}}(\mathbf{P}\mathbf{X}_0\mathbf{P}^{\mathrm{T}})^{-1}(\mathbf{c} - \mathbf{P}\mathbf{x}_0). \tag{12}$$

If other constraints are violated by this solution, $\mathbf{c}$ and $\mathbf{P}$ can be enlarged to accommodate them and (12) can be used to get a solution that satisfies these constraints as well.

At first glance it appears that computing the correction using (12) should be computationally undemanding, as the dimension of the matrix $\mathbf{P}\mathbf{X}_0\mathbf{P}$ corresponds to the number of constraints being enforced. However, the matrix $\mathbf{X}_0$ can be considerably more difficult to evaluate than the unconstrained estimate $\mathbf{x}_0$. When the data to be assimilated have uncorrelated errors, which is often the case, $\mathbf{D}$ is diagonal and evaluating $\mathbf{X}_0$ is not such a problem; it can be computed sequentially, one observation at a time. If the data can be partitioned into two or more groups with no cross-group error correlations, the update can be computed in a group-by-group sequential fashion at considerable savings, as computational cost increases like the third power of the rank of the inverted matrix. Thus direct application of (12) is a useful algorithm for enforcing constraints as long as correlations of data errors pose no problems.

This same sort of covariance update is encountered with the Kalman filter, which uses information carried by the assimilated data to reappraise uncertainties. Such updating of the error covariances together with their dynamical transformation from time step to time step makes Kalman filtering expensive. An argument for not accounting for the changing nature of the covariance matrix, i.e., of using optimal interpolation rather than Kalman filtering, is that our knowledge of the error covariances is sufficiently poor that we are not likely to be able to say whether the prior or posterior estimates better describe the state of uncertainty. The same reasoning here suggests an algorithm for approximating the constrained solution, namely replacing $\mathbf{X}_0$ with $\mathbf{B}$ when using (12) to enforce the inequality constraints.

It is interesting to compare the constrained solution (12) with the unconstrained solution (2). Note the similarity of the form of the two expressions. The vector $\mathbf{c}$ is like the data $\mathbf{d}$; the unconstrained solution $\mathbf{x}_0$ is like the background $\mathbf{b}$; the posterior covariance matrix $\mathbf{X}_0$ is like the background-error covariance matrix $\mathbf{B}$; and the constraint-selector matrix $\mathbf{P}$ is like the data-selector matrix $\mathbf{H}$; but there is no counterpart to the data-error covariance matrix $\mathbf{D}$, because the constraints are enforced with no allowance for error. In this regard the enforced constraints can be categorized as *strong*, while the data provide *weak* constraints Sasaki (1970). If the data $\mathbf{d}$ were error free, then $\mathbf{D} = 0$, and (2) would have precisely the same form as (12). On the other hand, if the constraints were treated as weak and included in the definition of the objective function as a quadratic term with matrix $\mathbf{C}^{-1}$, then (12) would have the same form as (2); letting $\mathbf{C} = 0$ would recover the strong-constraint expression.

This comparison suggests an alternative algorithm for enforcing the inequality constraints. After computing the unconstrained solution and discovering which constraints have been violated and need to be enforced, the constraint data $\mathbf{c}$ can be melded with the empirical data into an expanded $\mathbf{d}$ vector. Similarly, the data-error covariance matrix $\mathbf{D}$ should be enlarged with zero values for the constraint errors, and $\mathbf{H}$ should be enlarged to reflect the constraints. With these changes (2) can be used to compute the optimal constrained updates to the original background $\mathbf{b}$. The advantage of this algorithm is that it requires only minimal change to existing computational codes for assimilating data without regard to constraints. All that is necessary is to accommodate the expansion of the vectors and matrices. The basic logic remains essentially unchanged, requiring only the addition of a check for violated constraints that can trigger the necessary expansion and recomputation. The disadvantage is that it involves enlarged matrices, but it avoids the expense of computing the posterior covariance matrix $\mathbf{X}_0$.

With either algorithm or the approximate algorithm, if the solution violates any additional inequality constraints, the procedure can be repeated. No effort has been made to check that iteration should converge, as the emphasis here is on the practicalities of obtaining a useful solution to a poorly posed problem rather than on attaining a precise solution. Still, as each iteration can be expected to be smaller than the last and in the opposite direction, the need for subsequent iterations is unlikely.

4. Examples

For a simple computational example, consider a layer from HYCOM that is near enough to the surface for its thickness to be prescribed in some regions but deep enough to be isopycnic in others. Suppose that during a model simulation the thicknesses for four adjacent grid cells are 3, 2, 2, and 1 (times the minimum allowed value), while hydrographic data suggest the thickness for the first cell should be twice the minimum, i.e., more like what the model has for cell-2 and cell-3. Assimilating this information should certainly reduce the thickness for cell-1. If the background error covariances decrease exponentially with an e-folding distance equal to the width of a grid cell, due to the positive between-cell correlations, assimilation should also reduce cell-2's layer thickness to a lesser extent. An even smaller reduction can be anticipated for cell-3. With cell-4's layer thickness already at its limit, correcting cell-1 can be expected cause the thickness for cell-4 to fall below its minimum, and require its range constraint to be enforced.

For this small example it is easy to exhibit the matrices explicitly. The background estimate is $\mathbf{b} = (3\,2\,2\,1)^{\mathrm{T}}$; the data vector is simply $\mathbf{d} = 2$; $\mathbf{H} = (1\,0\,0\,0)$; and $\mathbf{Hb} = 3$. If the error variance of the data is 0.01 that of the background,[6] then $\mathbf{D} = 0.01$ and

$$\mathbf{B} = \begin{pmatrix} 1.0000 & 0.6065 & 0.3679 & 0.2231 \\ 0.6065 & 1.0000 & 0.6065 & 0.3679 \\ 0.3679 & 0.6065 & 1.0000 & 0.6065 \\ 0.2231 & 0.3679 & 0.6065 & 1.0000 \end{pmatrix}, \tag{13}$$

so $\mathbf{HBH}^{\mathrm{T}} = 1.000$ and $\mathbf{BH}^{\mathrm{T}} = (1.0000 \quad 0.6065 \quad 0.3679 \quad 0.2231)^{\mathrm{T}}$. Substituting into (2) gives the preliminary solution: $\mathbf{x}_0 = (2.0099 \quad 1.3995 \quad 1.6358 \quad 0.7791)^{\mathrm{T}}$. Layer thickness for cell 4 is below the minimum allowed value of 1.0 (see Fig. 5).

The inequality constraint for cell 4 can be enforced using the first algorithm described in Section 3 above, i.e., direct application of (12). Because there is only one value to be assimilated, the posterior covariance matrix is easily evaluated[7]

$$\mathbf{X}_0 = \begin{pmatrix} 0.0099 & 0.0060 & 0.0036 & 0.0022 \\ 0.0060 & 0.6358 & 0.3856 & 0.2339 \\ 0.0036 & 0.3856 & 0.8660 & 0.5253 \\ 0.0022 & 0.2339 & 0.5253 & 0.9507 \end{pmatrix}. \tag{14}$$

The constraint vector is simply $\mathbf{c} = 1$; the constraint-selector matrix is $\mathbf{P} = (0\,0\,0\,1)$; $\mathbf{PX}_0\mathbf{P}^{\mathrm{T}} = 0.9507$; and $\mathbf{X}_0\mathbf{P}^{\mathrm{T}} = (0.0022 \quad 0.2339 \quad 0.5253 \quad 0.\ 9507)^{\mathrm{T}}$. Eq. (12) gives the layer thicknesses after the constraint has been enforced: $\mathbf{x} = (2.0104 \quad 1.4538 \quad 1.7578 \quad 1.0000)^{\mathrm{T}}$.

Fig. 5 illustrates the results for this example. Background values of layer thickness are indicated by the upper curve, and the empirical layer thickness is indicated by the circle. The lower curve indicates the results of assimilation without regard for minimal layer thickness. The minimum allowed thickness (same for all cells) is indicated by the thin horizontal line. As expected, assimilation without regard for the constraints has

[6] Difference between background and observation must be relatively large and the accuracy of the measurement must be sufficiently greater than that of its background counterpart to provoke a noticeable constraint violation.

[7] Compare the diagonal elements of (13) and (14) and note the increased accuracy for cell 1 and the more moderate increases for the other cells, reflecting the high accuracy of the observation.

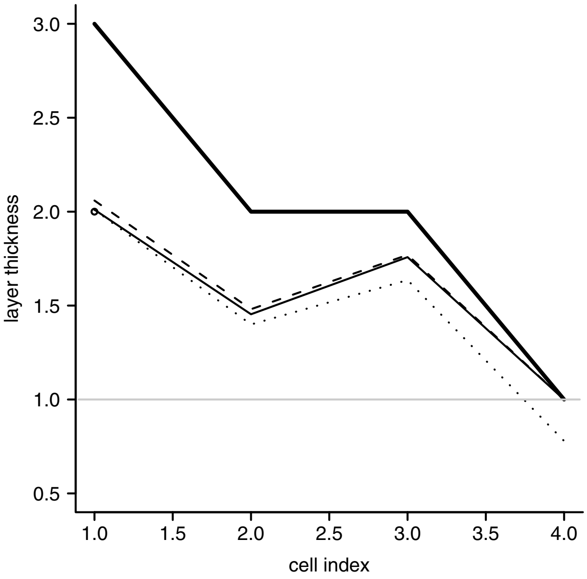

Fig. 5. Upper curve (heavy line) indicates background estimates and circle indicates observation to be assimilated. Lower curve (dotted line) indicates results after assimilation without accounting for inequality constraints. Horizontal line indicates minimum allowed layer thickness. Solid intermediate curve indicates results after enforcing constraint for cell-4 using the primary and alternate algorithms described above. Dashed curve indicates the results for the approximate algorithm.

resulted in cell-4 having a smaller than allowed thickness. And just as positive between-cell correlations propagated the reduction of thickness from cell-1 to the other cells, they should propagate the inflation of cell-4's thickness needed to satisfy the constraint back toward cell-1. This is illustrated by the solid intermediate curve, with the tiny adjustment for cell-1 too small to be seen on the scale of the figure.

The inequality constraint for cell 4 can also be enforced using the alternative algorithm described in Section 3. The data vector is expanded to include the constraining value: $\mathbf{d} = (2\ 1)^{\mathrm{T}}$. The data-error covariance matrix has zero entries for the constraint:

$$\mathbf{D} = \begin{pmatrix} 0.01 & 0 \\ 0 & 0 \end{pmatrix}; \tag{15}$$

and $\mathbf{H}$ is expanded to handle the constraint for cell-4 in addition to the observation for cell-1:

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{16}$$

With these changes to treat the constraint as data without error, (2) gives exactly the same result as found using the first algorithm.

The dashed curve on Fig. 5 is the solution obtained using the approximate algorithm, i.e., the first algorithm with $\mathbf{X}_0$ replaced by $\mathbf{B}$. Because $\mathbf{B}$ does not value the assimilated information as highly as $\mathbf{X}_0$ does, the original decreases in thickness to comply with the observation are re-inflated a bit too much by the constraint enforcement. Still, for this example the differences from the exact solutions are quite minor, and given its ease of use and the inadequacies of error estimates, this approximate algorithm is certainly worth consideration.

To explore the question of whether proceeding with the assumption that all constraint violations found during the first pass should be enforced without worrying whether a better solution might be obtained when only a subset is enforced, consider the situation where the hydrographic data suggest cell-1's thickness should be less than cell-2's. Fig. 6 shows that assimilation without regard for constraints causes both cell-2 and cell-4 to become too thin. The dashed curve showing the results where both are enforced lies almost on top of the solid curve representing the optimal solution with only the constraint for cell-4 enforced. Both results agree

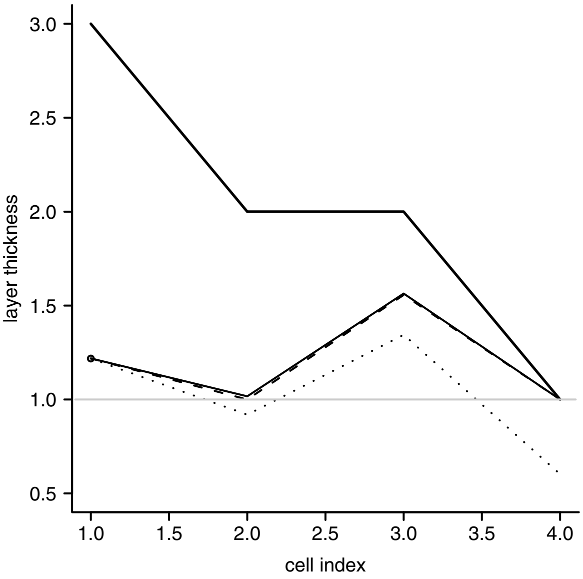

Fig. 6. Upper curve (heavy line) indicates background estimates and circle indicates observation to be assimilated. Lower curve (dotted line) indicates results after assimilation without accounting for inequality constraints. Horizontal line indicates minimum allowed layer thickness. Solid intermediate curve indicates results after enforcing only constraint for cell-4. Dashed curve indicates the results after enforcing constraints for both cell-2 and cell-4.

that cell-4 should have minimum thickness and both have essentially the same results for cell-1 and cell-3. The largest difference is for cell-2, which is allowed to have a thickness that is only slightly larger than minimum. This example supports the assumption that the expense of seeking the mathematical minimum of J is not justified, especially when the vagaries of the error statistics are considered.

The simple computational examples presented here illustrate the nature of the secondary corrections. Enforcing these constraints can be expected to be consequential only in the near neighborhood of the violation, and the secondary corrections are considerably smaller than those due to the initial data assimilation. The question remains concerning the extent to which these simple examples characterize what might be encountered when assimilating data into HYCOM with all its complications. The answer lies in recognizing that the primary difference is in the size of the variational problem and the number of constraints. In the vicinity of any particular constraint violation the situation is quite similar to what has been illustrated here. The scale of the spatial correlations together with the relative accuracy of the observations determine the magnitude and extent of the corrections. Here the data were given substantial influence in order to dramatize the problem; when the background estimate is considered to be more accurate, the problem diminishes.

5. Conclusion

The algorithms discussed in Section 3 and illustrated in Section 4 are presented as part of a two-step approach to incorporating inequalities into the standard variational formalism for data assimilation. The first step is to assimilate data without regard for the constraints and see which constraints have been violated; the second step is to enforce *all* identified violations explicitly. While there are situations where only a subset of violations need to be enforced explicitly, as their enforcement guarantees the enforcement of the others, this assumption avoids the computational burden of determining the members of that minimal subset. If that subset were known, the algorithms presented here could be used to effect these secondary enforcements as well as the adjustments of other variables to the presence of the constraints.

The emphasis of this paper is on computational efficiency. Because the statistics of background errors are quite poorly known in practice, it is easy to argue that there is a degree of imprecision in what constitutes an optimal solution. Consequently, there is a point of diminishing returns where additional precision in solving a poorly formulated problem is not warranted. If a precise solution to the mathematical problem is desired and computational cost is not an issue, techniques of nonlinear programming can be used and no assumption would be needed concerning which constraints to enforce. However, the underlying indeterminacy of the mathematical problem, which is due to imprecise knowledge of error statistics, i.e., to the arbitrariness of the objective function defining the problem, argues against using such techniques for data assimilation in the presence of inequality constraints. The approach suggested here offers a more economical alternative.

The analysis presented here shows that, when the constraints to be enforced are known, they can be treated as error-free data and used to compute the secondary corrections. The first algorithm computes a correction to the unconstrained solution, whereas the second combines the constraints with the original data to get a revised solution. Which to use depends entirely on ease of implementation within existing software for assimilating data. An approximation to the first algorithm, which for computational economy computes the correction without bothering to account for the changes in error covariances implied by the first step, provides a third choice. If any of these produce additional violations, the process can be iterated. As each iteration can be expected to yield smaller corrections than the proceeding, the practical issue reduces to whether to bother at all with the secondary corrections or simply to correct the violations resulting from the unconstrained violation without bothering with secondary corrections and be done.

As the motivation for this work has been respecting the nature of HYCOM's layers while assimilating data, it is appropriate to conclude with a few comments about the practical issues that might be encountered. First, HYCOM presents other constraints that should be addressed. In particular, to avoid stimulating an adverse barotropic response, it is best to confine the influence of the data to the baroclinic modes; this requires that for each grid cell the sum of layer thicknesses for the entire water column be the same after assimilating data as before. Such equality constraints can also be accommodated into the formalism with Lagrange multipliers as described in Section 3, but as the focus here is on inequality constraints, that issue was not pursued. As all layers might be used in only in the deepest regions and some might be collapsed into zero thickness when the bottom is too shallow to allow for water of their target densities, there is the question of deciding dynamically which layers should be active as layers slosh up and down the slope; such considerations can be regarded as additional constraints. In the work described by Thacker et al. (2004) the simple expedient of enforcing constraints while ignoring all secondary corrections implied by error correlations caused no adverse problems. The approach presented here behaves about the same except that fields are a bit smoother in the vicinity of the enforced constraints. An issue touched on above is how uncertainties in the background state really should be characterized. Here truncated normal distributions were used for mathematical convenience, as little is known about the statistics of layer-thickness errors. As the error distributions control the assimilative corrections, much work is needed in this area. Finally, the algorithms discussed here are appropriate when the assimilation system is implemented in terms of formulae like Eq. (2). If it is implemented to seek the minimum of the objective function J of (1) using an algorithm like conjugate-gradient descent, the constraints can be enforced at each iteration with the idea that the secondary corrections will be picked up by subsequent iterations, but the impact on the performance of the descent algorithm should be investigated.

Acknowledgements

This work was supported by the National Oceanographic Partnership Program and by the Atlantic Oceanographic and Meteorological Laboratory.

References

Behringer, D.W., Ji, M., Leetmaa, A., 1998. An improved coupled model for ENSO prediction and implications for ocean initialization. Part I: The ocean data assimilation system. Mon. Wea. Rev. 126, 1013–1021.

Bleck, R., 2002. An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates. Ocean Modell. 37, 55–88.

Bretherton, F.E., Davis, R.E., Fandry, C.B., 1976. A technique for objective analysis and design of oceanographic experiments applied to mode-73. Deep-Sea Res. 23, 559–582.

Carton, J.A., Chepurin, G., Cao, X., Giese, B., 2000. A simple ocean data assimilation analysis of the global upper ocean 1950–1995. J. Phys. Oceanogr. 30, 294–309.

Cohn, S.E., 1997. An introduction to estimation theory. J. Meteor. Soc. Jpn. 75, 257–288.

Daley, R., 1991. Atmospheric Data Analysis. Cambridge University Press, Cambridge.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis. John Wiley and Sons, New York.

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

Gandin, L.S., 1963. Objective Analysis of Meteorological Fields. Israel Program for Scientific Translations, Jerusalem, translated from the Russian.

Gelb, A. (Ed.), 1974. Applied Optimal Estimation. Springer-Verlag, Cambridge, MA.

Gill, P.E., Murray, W., Wright, M.H., 1981. Practical Optimization. Academic Press, London.

Gnanadesikan, A., 1999. Numerical issues for coupling biological models with isopycnal mixing schemes. Ocean Modell. 1, 15–883.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Trans. Am. ASME, Ser. D, J. Basic Eng. 82, 35–45.

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

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. Quart. J. Roy. Meteor. Soc. 112, 1177–1194.

Marotzke, J., Giering, R., Zhang, Q.K., Stammer, D., Hill, C.N., Lee, T., 1999. Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity. J. Geophys. Res. 104, 29529–29548.

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

Smolarkiewicz, K.P., 1984. A fully multidimensional positive definite advection transport algorithm with small implicit diffusion. J. Comput. Phys. 54, 325–362.

Thacker, W.C., 1989. The role of the Hessian matrix in fitting models to measurements. J. Geophys. Res. 94, 6177–6196.

Thacker, W.C., Esenkov, O.E., 2002. Assimilating XBT data into HYCOM. J. Atmos. Oceanic Technol. 19 (5), 709–724.

Thacker, W.C., Long, R.B., 1988. Fitting dynamics to data. J. Geophys. Res. 93, 1227–1240.

Thacker, W.C., Lee, S.-K., Halliwell Jr., G.R., 2004. Assimilating 20 years of Atlantic XBT data into HYCOM: A first look. Ocean Modell. 7, 183–210.